

(12)特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局(43) 国際公開日
2003 年 10 月 16 日 (16.10.2003)

PCT

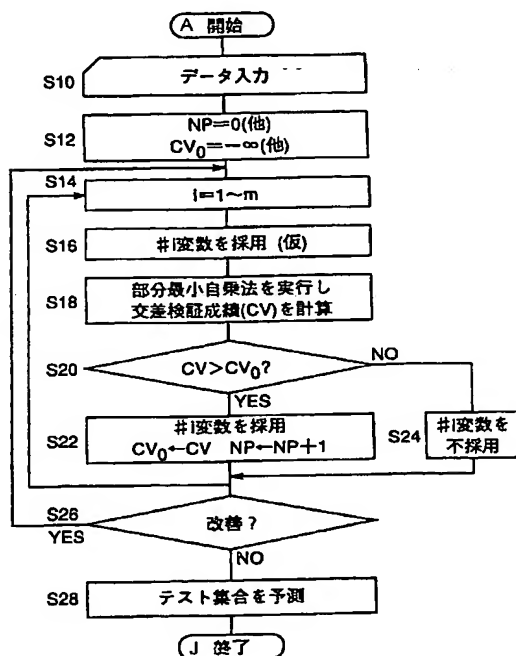
(10) 国際公開番号
WO 03/085548 A1

- (51) 国際特許分類⁷: G06F 17/18, 17/30, (71) 出願人 (米国を除く全ての指定国について): 石原産業株式会社 (ISHIHARA SANGYO KAISHA, LTD.)
C12N 15/00, C12Q 1/68, G01N 33/574 [JP/JP]; 〒550-0002 大阪府 大阪市 西区江戸堀一丁目 3 番 1 5 号 Osaka (JP).
- (21) 国際出願番号: PCT/JP03/04059
- (22) 国際出願日: 2003 年 3 月 31 日 (31.03.2003) (72) 発明者; および
- (25) 国際出願の言語: 日本語 (75) 発明者/出願人 (米国についてのみ): 石川 俊夫 (ISHIKAWA, Toshio) [JP/JP]; 〒525-0025 滋賀県 草津市 西浜川二丁目 3 番 1 号 石原産業株式会社 中央研究所内 Shiga (JP). 久米 隆志 (KUME, Takashi) [JP/JP]; 〒525-0025 滋賀県 草津市 西浜川二丁目 3 番 1 号 石原産業株式会社 中央研究所内 Shiga (JP).
- (26) 国際公開の言語: 日本語
- (30) 優先権データ:
特願2002-102743 2002 年 4 月 4 日 (04.04.2002) JP
特願2002-352645 2002 年 12 月 4 日 (04.12.2002) JP

[続葉有]

(54) Title: APPARATUS AND METHOD FOR ANALYZING DATA

(54) 発明の名称: データ解析装置および方法



A...START
 S10...DATA INPUT
 S12...NP=0 (OTHERS) CV₀=-∞(OTHERS)
 S16...(TENTATIVE) EMPLOYMENT OF VARIABLE #i
 S18...CALCULATION OF CROSS VERIFICATION (CV) BY
 PARTIAL LEAST SQUARES METHOD
 S22...EMPLOYMENT OF #i CV₀ CV NP NP+1
 S24...NO EMPLOYMENT OF #i
 S26...IMPROVED?
 S28...ESTIMATION OF TEST MASS
 J...FINISH

(57) Abstract: In data analysis for determining a correlation model among a condition of a living body and the expression doses of plural genes and/or the amount of an intracellular substance, an explaining variable contained in the data is selected in a data mass wherein the condition of the living body or a change in the living body probabilistically occurring with the passage of time is regarded as a target variable while the expression doses of plural genes and/or the amount of an intracellular substance are regarded as explaining variables. Then the cross verification of a correlation model containing the thus selected explaining variable and target variable is calculated and the results are evaluated. The selection of the explaining variable, the calculation of cross verification and the evaluation of the results are repeated until the cross verification is not improved any more to thereby determine a partial least squares method model. Thus, an effective data processing method for multivariable gene expression data is provided.

(57) 要約: 生体の状態と複数の遺伝子発現の量および/または細胞内物質の量との相関モデルを決定するデータ解析において、生体の状態または時間とともに確率的に発生する生体の状態の変化を目的変数とし、複数の遺伝子発現の量および/または細胞内物質の量を説明変数とするデータの集合において、データに含まれる説明変数を選択し、選択された説明変数と目的変数とを含む相関モデルについて交差検証成績を計算し、その結果を評価判定する。ここで、交差検証成績が改善しなくなるまで、説明変数の選択、交差検証成績の計算、その結果の評価判定を行い、部分最小自乗法モデルを決定する。これにより、多変量の遺伝子発現情報の効果的な情報処理を提供する。



(74) 代理人: 青山 葆, 外(AOYAMA, Tamotsu et al.); 〒540-0001 大阪府 大阪市 中央区 城見 1 丁目 3 番 7 号
I M P ビル 青山特許事務所 Osaka (JP).

(81) 指定国 (国内): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NI, NO, NZ, OM, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) 指定国 (広域): ARIPO 特許 (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), ユーラシア特許 (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), ヨーロッパ特許 (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, RO, SE, SI, SK, TR), OAPI 特許 (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

添付公開書類:

— 国際調査報告書

2 文字コード及び他の略語については、定期発行される各 PCT ガゼットの巻頭に掲載されている「コードと略語のガイダンスノート」を参照。

明 細 書

データ解析装置および方法

5 技術分野

本発明は、生体の状態と遺伝子発現の量および／または細胞内物質の量との多変量解析処理並びそれを基に可能となる測定機材、検定方法などに関するものである。

10 背景技術

2000年6月のヒトゲノムの解読宣言以降、ゲノムに書かれた遺伝情報がどのように発現して機能しているかのを解明するポストゲノム時代に突入したと言われている。ヒトゲノム計画の進展の中で、ゲノム発現状態を測定する方法論も進展してきた。トランスクリプトーム(mRNA)測定手段としてオリゴヌクレオチドアレイやマイクロチップが知られている。またプロテオーム(蛋白質)測定手段として、以前からある2次元電気泳動に加えて、最近では質量分析の方法が進歩してきた。また抗体チップなどの先進の技術も注目されている。これらの測定技術は、生体の状態パラメータを短時間に一挙に測定できることがそれまでの技術と比較して画期的であるといえる。

20 遺伝子発現状態を効率的に測定する技術として次のものがあげられる。トランスクリプトーム(mRNAの総体)を特定するものとして、基盤に複数種のDNAを担持し、それに相補的なmRNAを検出するDNAチップが知られている。代表的なDNAチップには、遺伝子チップやDNAマイクロアレイがある。また、プロテオーム(蛋白質の総体)を特定するものには、2次元電気泳動、抗体チップ、質量スペクトルを用いるものがある。またメタボローム(代謝中間体を含めた代謝産物の総体)を測定する手法も質量分析などによって試みられており、進展が見られる。

25 生体内の細胞の状態は遺伝子産物の発現によってよく記述されるため、従来の診断マーカーでは情報が不足している場面でも、精度のより高い診断が可能にな

るという期待も出てきている。たとえば、次のような研究があげられる。

P. O. Brownらは、DNAチップによってリンパ腫患者の細胞のトランスクリプトームを測定し、クラスター解析によって悪性と良性のリンパ腫（DLBCL）を別クラスターに分離した（Nature 403(3), 503-11 (2000)）。しかし、これは因果関係（相関関係）のモデルを得る方法ではなく、どの遺伝子がどの程度重要かを判断できない。

A. Alaiyaらは、2次元電気泳動によって子宮がん患者40人の細胞のプロテオームを測定し、うち22人のデータから部分最小自乗法診断モデルを構築し、悪性度を説明した（Int. J. Cancer, 86, 731-36 (2000); Electrophoresis, 21, 1210-17 (2000); 国際公開 WO 00/70340）。その際、全変数モデルにおいて153変数からloadingの大きな170変数に限定することによって交差検証成績がよくなり（ $Q^2 = 0.84$ ）、残り18患者の深刻度（3段階）を11/18の比率で正答した。交差検証法がモデル構築の際の指標になるという考えが表明されている。しかし、この方法では、loadingを得る際にまず全変数モデルが成立しなければならない。また、それ以外の変数選択手法が考案されていない。

J. Khanらは、DNAチップによって小児がん患者の細胞を測定し、ニューラルネットワークによって悪性度を説明した（Nature Medicine, 7(6), 673-79 (2001)）。小児がん（SRBCT）患者88人のトランスクリプトーム（6567遺伝子）を測定し、うち63人のデータから主成分分析によって10次元に圧縮し、次に、人工ニューラルネットワーク診断モデルを構築した。ここで、影響力のある上位遺伝子を交差検証法によって絞り込み、96遺伝子で最良の成績(100%)を得た。このモデルで残り25人を予測し、93～100%の結果を得た。しかし、この方法でも、影響力を得る際にまず全変数モデルが成立しなければならない。またそれ以外の変数選択手法が考案されていない。10次元のような少ない変数の場合を扱えるが、変数の数が膨大な場合には適用できない。

また、最近になってDNAチップの解析に部分最小自乗法を用いる研究がD. M. RockeとD. V. Nguyenによって報告されるに至った（国際公開 WO 02/25405; Bioinformatics 18(1), 39-50 (2002); Bioinformatics 18(9), 1216-26 (2002); Bioinformatics 18(12), 1625-32 (2002)）。部分最小自乗法の潜在変数

を線型判別分析などの多変量解析の説明変数として用いた場合に良好な結果が得られることが報告されている。これは部分最小自乗法が次元圧縮とモデルフィットを同時に行なうことのできる方法であるために可能となったものである。報告に示された実施例では部分最小自乗法がDNAチップ情報のモデル構築方法として優れたものであることが示されている。しかし報告においては重要な遺伝子発現量を選抜する手段としての最小自乗法の適用については触れられておらず、事前の前処理によって選択された説明変数を全て用いて解析が行なわれているという点において上述のA. Alaiyaらの研究と同様の課題を含んでいる。

従来の診断マーカーでは情報が不足している場面でも、遺伝子発現情報を活用することで、より精度（解像度）の高い診断が可能になるという期待も出てきている。遺伝子発現状態の測定結果は、膨大な情報量が得られることが従来にはなかった特徴であり、逆に情報量が多いために、効果的なデータ処理なくしてデータの活用はありえない。したがって、有用な知識を獲得するためには効果的な情報処理が欠かせない。前に説明したように、現状ではクラスター解析を中心とする方法が用いられているが、主成分分析などの方法も採用されている。クラスター解析や主成分分析は、教師付学習方法ではないため、病状の因果関係（相関関係）のモデルを得ることはできない。すなわち、どの遺伝子がどの程度重要かを解析結果から得ることができないのが難点である。一方、部分最小自乗法は次元圧縮とモデルフィットを同時に行なう強力な多変量解析手法であるが、変数の数が膨大になった場合にしばしば有意な結果が得られない事態に直面する。したがって、膨大な遺伝子発現情報などから有用な知識を獲得できるような効果的な情報処理が望まれている。また、そのような情報処理の結果を基にした効率的な測定機材、検定処理などが期待されている。

発明の開示

（発明が解決しようとする技術的課題）

この発明の目的は、多変量の遺伝子発現情報、細胞内物質情報の効果的な情報処理を提供することである。

また、この発明の目的は、効率的な検定処理を提供することである。

(その解決方法)

本発明に係るデータ解析装置は、生体の状態または時間とともに確率的に発生する生体の状態の変化を目的変数とし、複数の遺伝子発現の量および／または細胞内物質の量を説明変数とする相関モデルを決定するデータ解析装置であって、

5 生体の状態或いはそれを導出するデータまたは時間とともに確率的に発生する生体の状態の変化に関するデータと、複数の遺伝子発現の量および／または細胞内物質の量からなるサンプルの集合を入力する入力手段と、(1)説明変数を選択する選択手段と、(2)部分最小自乗法を実行して交差検証成績を計算する計算手段または上記生体の状態の変化に関するデータにカプラン・マイヤー法又はカトラ

10 ー・エデラー法による生命表を適用して変化の発生しなかったものの確率を計算して得られた確率を、仮定した分布に基づいた変換または仮定を前提としない変換をし、該変換結果を目的変数とする部分最小自乗法を実行して交差検証成績を計算する計算手段と、(3)上記(2)の計算手段の結果を評価し、説明変数の採用、不採用を判定する評価判定手段とを有し、(4)上記(1)の選択手段と上記(2)

15 の計算手段と上記(3)の評価判定手段とを実行して部分最小自乗法モデルの少なくとも交差検証成績を独立変数として持つ関数を改善し続けて部分最小自乗法モデルを決定する決定手段とからなる。選択手段は、たとえば、説明変数を逐次取捨選択したり、遺伝的アルゴリズムを用いて説明変数を選択する。計算手段は、たとえば、1個のサンプルを逐次除外したり、複数のサンプルを逐次除外して部

20 分最小自乗法を実行して交差検証成績を計算する。評価判定手段は、たとえば、計算手段の結果から、各計算において除外したサンプルの遺伝子発現から予測される生体の状態を示す目的変数値と、前記除外したサンプルの生体の状態を示す目的変数値との誤差の代表値を求め、当該誤差の代表値が小さくなった場合に、その交差検証成績が改善されたと判定し、説明変数を取捨選択しながら交差検証

25 成績の評価判定を繰り返す。あるいは交差検証成績ではなく、少なくとも部分最小自乗法モデルの交差検証成績を独立変数として持つ関数が改善するかどうかを評価判定の基準として用いることもできる。決定手段は、たとえば、選択手段と計算手段と評価判定手段とを繰り返し実行して部分最小自乗法モデルの交差検証成績を改善し続けて部分最小自乗法モデルを決定する。また、選択手段と計算手

段とを複数のコンピュータで実行させることもできる。こうして、相関モデルを構成するとき、交差検証成績を基準に最適化させることにより説明変数を取捨選択し、説明変数の次元を減らして良好なモデルを得る。

上述の、仮定した分布に基づいた変換または仮定を前提としない変換は、生体の状態の変化の確率が説明変数の多項式で解析できるようにするために行なうものである。分布を仮定した場合には、確率を対数変換後に負の数にしたものを状態の変化を観測した時間で割るという変換、確率を対数変換後に負の数にしたものをさらに対数にしたものを状態の変化を観測した時間で割るという変換、または確率を1より減じたものをプロビット変換したものを計算して状態の変化を観測した時間で割るという変換などが考えられる。一方、分布を仮定しない場合にはロジット変換といった方法が考えられる。変換の方法は分布にどのような仮定が成り立つかどうかあるいはなりたたないかどうかを判断することにより、それぞれの場合に応じて適切に選ぶことができる。少なくとも部分最小自乗法モデルの交差検証成績を独立変数として持つ関数としては、たとえば、前記誤差の代表値と選抜された説明変数の数の関数が考えられ、あるいはその他の独立変数を含むものであってよい。望ましくは、関数は誤差の代表値の単調減少関数であり、説明変数の数の単調減少関数である。計算量を増やさないためには簡単に計算できる関数が望ましい。具体的には $-\text{PRESS} \times \alpha^{N/P}$ という関数が考えられる。ここでPRESSは予測残差自乗和であり、 N/P は採用された説明変数の数であり、 α は1または1より大きい実数である。また、 $-\text{PRESS} \times (N/P + \beta)^{-\gamma}$ や $-\text{PRESS} \times (\beta - N/P)^{-\gamma}$ なる関数も考えられる。ここで、 γ は正の実数である。

説明変数の個数を少なくすると、通常の統計的手法または多変量解析手法が適用可能になる。本発明では部分最小自乗法を用いて選抜された説明変数を統計手法又は多変量解析手法の説明変数として、より良好なモデルを得る。或いは選抜された説明変数を用いた部分最小自乗法モデルの潜在変数を統計手法又は多変量解析手法の説明変数として、より良好なモデルを得る。ここで潜在変数とは、部分最小自乗法において通常用いられているものであって、目的変数(Y_{ij})と説明変数(X_{ij})の背後に共通する次元数の少ない潜在変数(T_{ik})を抽出することが部分

最小自乗法の次元圧縮であり、モデルフィットである。

$$Y_{il} = \sum Q_{kl} \times T_{ik} + F_{il}$$

$$X_{ij} = \sum P_{kj} \times T_{ik} + E_{ij}$$

(iはサンプル番号、lは目的変数番号、jは説明変数番号、kは潜在変数番号、

F, Eは残差)

また、統計的手法又は多変量解析手法としては、重回帰分析法、線型判別分析法、適応最小自乗法、ロジスティック回帰分析法、比例ハザード解析法、マハラノビス距離を用いる判別分析法、kNN法、人工ニューラルネットワークなどが挙げられる。

本発明者等は、また、 Q^2 やPRESS値などの交差検証成績に加えて、説明変数の個数を第2の独立変数として含む関数を最適化することで選抜される説明変数を任意に絞り込むことができることを新たに見出した。通常の統計的手法や多変量解析手法では、抽出される説明変数の個数NPの望ましい範囲がサンプル数との兼ね合いで決まっている場合がある。そのような場合、関数を、目的とする選抜数によって任意に変更できる。関数形をたとえば $\text{PRESS} \times \alpha^{NP}$ とした場合、説明変数の個数を数個から数十個に絞り込むためには通常は定数alphaとして1.0～3.0の値が望ましい。より望ましくは、alphaは1.0～2.0の値となる。他の関数形 $f(\text{PRESS}, NP)$ であっても、実際に選抜される説明変数の数MPおよびその時のPRESS値PRESS_MPの周辺で、 $f(\text{PRESS_MP} \div \alpha, MP+1) \approx f(\text{PRESS_MP}, MP)$ となるような関数は、変数選抜という点では同様の効果を持つ場合がある。こうして、適当な関数形を用いることにより、望ましい範囲の個数NPの説明変数を選抜できる。このようにして、交差検証成績を用いて決定されたモデルに採用されている説明変数をさらに絞り込むと、統計的手法又は多変量解析手法によるモデルを構築できる。したがって、その性質が十分解明されている統計的手法又は多変量解析手法を採用して解析を加えることができる。

また、目的変数として、時間とともに確率的に発生する生体の状態の変化から導出された量を用いて、時間とともに確率的に発生する生体の状態の変化と複数の遺伝子発現の量および／または細胞内物質の量との相関モデルを決定できる。

「時間とともに確率的に発生する生体の状態の変化」とはたとえば生存時間であ

る。ここで、前述の部分最小自乗法に、カプラン・マイヤー法又はカトラー・エデラー法と、ロジット(logit)変換とを組み合わせる。部分最小自乗法での目的変数は、時間とともに確率的に発生する生体の状態の変化に関するデータにカプラン・マイヤー法又はカトラー・エデラー法による生命表を適用して変化の発生しなかったものの確率を計算し、これをロジット変換した値である。ロジット(logit)値とは、分類分けされたデータの、ある分類の割合(確率) P を基に、次式 $\text{logit} = \log \{P/(1-P)\}$ にて計算される値である。ロジット値を目的変数とする部分最小自乗法を実行して交差検証成績を計算する。こうして、先に説明したのと同様に、部分最小自乗法の交差検証成績を考慮した説明変数の抽出を行って、生存時間解析を行える。

説明変数の個数を少なくすると、通常の統計的手法または多変量解析手法が適用可能になる。そこで、決定されたモデルに採用されている説明変数又はその潜在変数を用い、時間とともに確率的に発生する生体の状態の変化を説明する統計的手法又は多変量解析手法によるモデルを構築する。たとえば、ロジット値を目的変数として求めた説明変数を用いて、他の統計的手法又は多変量解析手法(たとえば比例ハザード法や、パラメトリックな分布にあてはめた回帰分析法)を行なうことによって、より良好なモデルを得ることができる。比例ハザード法とは、Coxによって考案された方法であり、生存率の解析に時間を考慮し、かつ、多変量を扱える。比例ハザード法では、観測されている個々ごとにハザード値と呼ばれる生存率を左右する値があり、それを導く関数がある(モデルが仮定されている)として解析される。カプラン・マイヤー法は、集団全体または群ごとの生存率の推移を示す。また、パラメトリックな分布とは、ガウスが提案した正規分布から計算された確率分布のことであり、生存時間解析では指数分布、ワイブル分析、対数正規分布が用いられる。指数分布などへの当て嵌めで、数式中に多項式があり、前述の部分最小自乗法の交差検証成績を考慮した説明変数の抽出が適用される。

入力手段で説明変数として入力される複数の遺伝子の発現量および/または細胞内物質の量とは、必ずしも物質の絶対的な濃度の測定値に限定されるものではなく、加工計算された値、相対的な値、間接的に物質量を表す量などでもよい。

たとえば、質量スペクトルで蛋白質の発現量を測定することができることを応用して、生体の状態を表わす目的変数と、質量スペクトルとを直接関係づける相関モデルを構築することができる。またAffymetrix社タイプのDNAチップ(ジェンチップ)では、単一のスポットが単一の遺伝子発現を特定するとは限らず、複数個のスポットが集まってはじめて単一の遺伝子発現を特定することもある。ここでもまた、各スポットの測定量を説明変数として、直接、生体の状態を説明する相関モデルを得ることができる。更には、タンパク質の電気泳動パターン of 各ピークは単一のタンパク質に帰属できず、複数個のタンパク質の重ねあわせであることも多い。このような場合にも生体の状態を説明する説明変数として各ピーク強度を用いることができる。このことは、上述のAlaiyaらは子宮癌の診断の説明変数として電気泳動パターンのピーク強度を採用していることから明らかである。前述のようにポストシーケンス時代のトランスクリプトーム解析、プロテオーム解析、メタボローム解析という研究分野では、生体(細胞)内の物質を総体として把握することから出発することを特徴とする実験的アプローチが注目されている。ひとつひとつの物質の絶対的定量は必須事項ではなく、これらの実験方法によって定量される物質の量を直接、間接に表現する測定値やその加工計算値が、生体の状態を説明する説明変数と成り得る。また以上の物質量を表現する説明変数以外に、場合によっては問診データなどの他の説明変数を追加すると、さらに有効な解析結果が得られる場合もある。

本発明に係るデータ解析方法は、生体の状態または時間とともに確率的に発生する生体の状態の変化を目的変数とし、複数の遺伝子発現の量および/または細胞内物質の量を説明変数とする相関モデルを決定するデータ解析方法であって、生体の状態或いはそれを導出するデータまたは時間とともに確率的に発生する生体の状態の変化に関するデータと、複数の遺伝子発現の量および/または細胞内物質の量からなるサンプルの集合を入力する入力ステップと、(1)説明変数を選択する選択ステップと、(2)部分最小自乗法を実行して交差検証成績を計算する計算ステップまたは前記生体の状態の変化に関するデータにカプラン・マイヤー法又はカトラー・エデラー法による生命表を適用して変化の発生しなかったものの確率を計算して得られた確率を、仮定した分布に基づいた変換または仮定を前

提としない変換をし、該変換結果を目的変数とする部分最小自乗法を実行して交差検証成績を計算する計算ステップと、(3)前記(2)の計算ステップの結果を評価し、説明変数の採用、不採用を判定する評価判定ステップとを有し、(4)前記(1)の選択ステップと前記(2)の計算ステップと前記(3)の評価判定ステップとを実行して部分最小自乗法モデルの少なくとも交差検証成績を独立変数として持つ関数を改善し続けて部分最小自乗法モデルを決定する決定ステップとからなる。

このデータ解析方法において、選択ステップは、たとえば、説明変数を逐次取捨選択したり、遺伝的アルゴリズムを用いて説明変数を選択する。計算ステップは、たとえば、1個のサンプルを逐次除外したり、複数のサンプルを逐次除外して部分最小自乗法を実行して交差検証成績を計算する。評価判定ステップは、たとえば、計算ステップの結果から、各計算において除外したサンプルの遺伝子発現から予測される生体の状態を示す目的変数値と、前記除外したサンプルの生体の状態を示す目的変数値との誤差の代表値を求め、当該誤差の代表値が小さくなった場合に、その交差検証成績が改善されたと判定し、説明変数を取捨選択しながら交差検証成績の評価判定を繰り返す。決定ステップは、たとえば、選択ステップと計算ステップと評価判定ステップとを繰り返し実行して部分最小自乗法モデルの交差検証成績を改善し続けて部分最小自乗法モデルを決定する。また、選択ステップと計算ステップとを複数のコンピュータで実行させることもできる。

本発明に係るデータ解析プログラムは、生体の状態または時間とともに確率的に発生する生体の状態の変化を目的変数とし、複数の遺伝子発現の量および／または細胞内物質の量を説明変数とする相関モデルを決定する、コンピュータにより実行されるデータ解析プログラムであって、生体の状態或いはそれを導出するデータまたは時間とともに確率的に発生する生体の状態の変化に関するデータと、複数の遺伝子発現の量および／または細胞内物質の量からなるサンプルの集合を入力する入力ステップと、(1)説明変数を選択する選択ステップと、(2)部分最小自乗法を実行して交差検証成績を計算する計算ステップまたは前記生体の状態の変化に関するデータにカプラン・マイヤー法又はカトラー・エデラー法による生命表を適用して変化の発生しなかったものの確率を計算して得られた確率を、

仮定した分布に基づいた変換または仮定を前提としない変換をし、該変換結果を目的変数とする部分最小自乗法を実行して交差検証成績を計算する計算ステップと、(3)前記(2)の計算ステップの結果を評価し、説明変数の採用、不採用を判定する評価判定ステップとを有し、(4)前記(1)の選択ステップと前記(2)の計算ステップと前記(3)の評価判定ステップとを実行して部分最小自乗法モデルの少なくとも交差検証成績を独立変数として持つ関数を改善し続けて部分最小自乗法モデルを決定する決定ステップとからなる。

このデータ解析プログラムにおいて、選択ステップは、たとえば、説明変数を逐次取捨選択したり、遺伝的アルゴリズムを用いて説明変数を選択する。計算ステップは、たとえば、1個のサンプルを逐次除外したり、複数のサンプルを逐次除外して部分最小自乗法を実行して交差検証成績を計算する。評価判定ステップは、たとえば、計算ステップの結果から、各計算において除外したサンプルの遺伝子発現から予測される生体の状態を示す目的変数値と、前記除外したサンプルの生体の状態を示す目的変数値との誤差の代表値を求め、少なくとも当該誤差の代表値を独立変数として持つ関数である当該誤差の代表値の単調減少関数の値が小さくなった場合に、その交差検証成績が改善されたと判定し、説明変数を取捨選択しながら交差検証成績の評価判定を繰り返す。決定ステップは、たとえば、選択ステップと計算ステップと評価判定ステップとを繰り返し実行して少なくとも部分最小自乗法モデルの交差検証成績を独立変数として持つ関数を改善し続けて部分最小自乗法モデルを決定する。また、選択ステップと計算ステップとを複数のコンピュータで実行させることもできる。さらには、前記の説明変数の選択において、たとえば、初期状態では説明変数を全く含まないか、或いは、初期状態では全説明変数を含むこともできる。

前記のデータ解析プログラムにおいて、上記の生体の状態は、たとえば病気のタイプをあらわす測定値、病気の重篤度をあらわす測定値、病気のタイプをあらわす医療診断の結果、病気の重篤度をあらわす医療診断の結果、あるいはそれらを2次加工した数値である。例えば後の実施例で示すように、患者の生存時間を予測することは、QOL(quality of life:生活の質)を含めた治療計画や人生設計などを判断する上で重要な情報をもたらすものであり、社会的に価値のある診断

モデルを提供することができる。また癌の再発可能性を予測することは、QOLを考慮した治療計画を立案し、医師または当の患者が選択の判断をするうえで、貴重な情報をもたらすものである。

また、本発明は、決定された前記相関モデル及び予測対象のサンプルについて
5 当該モデルにおいて採用された説明変数を入力する入力手段と、入力された該説明変数に基づいて該サンプルの生体の状態を予測判定する予測判定手段からなるデータ解析装置、前記で決定された相関モデル及び予測対象のサンプルについて
10 当該モデルにおいて採用された説明変数を入力する入力ステップと、入力された該説明変数に基づいて該サンプルの生体の状態を予測判定する予測判定ステップからなるデータ解析方法及び前記で決定された相関モデル及び予測対象のサンプルについて当該モデルにおいて採用された説明変数を入力する入力ステップと、
入力された該説明変数に基づいて該サンプルの生体の状態を予測判定する予測判定ステップからなるデータ解析プログラムも包含する。

本発明に係るコンピュータにより読取可能な記録媒体は、上記のいずれかのプログラムを記録する。
15

本発明に係るびまん性大細胞型Bリンパ腫の重篤度検定用の細胞内物質測定機材および測定方法並びにびまん性大細胞型Bリンパ腫の重篤度検定方法は、実質的にジーンバンクアクセッション番号がU15085、M23452、X52479、U70426、H57330及びS69790からなる遺伝子群の発現を検出する。さらに、ジーンバンクア
20 クセッション番号がU03398、M65066、AK001546、BC003536、X00437、U12979、H96306、AA830781及びAA804793からなる群から選択される少なくとも一つの遺伝子の発現を検出してもよい。

また、本発明に係る乳癌の重篤度検定用の細胞内物質測定機材および測定方法並びに乳癌の重篤度検定方法は、実質的にジーンバンクアクセッション番号が
25 AA598572、AA703058及びAA453345からなる遺伝子産物を含む細胞内物質を検出する。さらに、ジーンバンクアクセッション番号がAA406242、H73335、W84753、N71160、AA054669、N32820及びR05667からなる群から選択される少なくとも一つの遺伝子産物を含む細胞内物質を検出してもよい。

また、本発明に係る乳癌の再発性検定用の細胞内物質測定機材および測定方法

並びに乳癌の再発性検定方法は、実質的にジーンバンクアクセッション番号が W84753、H08581、AA045730及びAI250654からなる遺伝子産物を含む細胞内物質を検出する。さらに、ジーンバンクアクセッション番号がAA448641、R78516、R05934、AA629838及びH53037からなる群から選択される少なくとも一つの遺伝子産物を含む細胞内物質を検出してもよい。

また、本発明に係る乳癌の再発性検定用の細胞内物質測定機材および測定方法並びに乳癌の再発性検定方法は、実質的にジーンバンクアクセッション番号が AA434397、T83209、N53427、N29639、AA485739、AA425861、H84871、T64312、T59518及びAA037488からなる遺伝子産物を含む細胞内物質を検出する。さらに、ジーンバンクアクセッション番号がAA406231の遺伝子産物を含む細胞内物質を検出してもよい。

また、本発明に係る乳癌の再発性検定用の細胞内物質測定機材および測定方法並びに乳癌の再発性検定方法は、実質的にジーンバンクアクセッション番号が H11482、T64312及びAA045340からなる遺伝子産物を含む細胞内物質を検出する。

細胞内物質測定機材としては、DNAマイクロアレイ、ジーンチップ、オリゴDNA型のDNAチップ、電気化学DNAチップ(ECAチップ)、繊維型DNAチップ、磁性ビーズDNAチップ(PSS)、糸巻きDNAチップ(PSS)、などのDNAチップ、マクロアレイ、抗体チップ、測定用試薬キットなどが挙げられる。また、上記の機材を適宜組み込んだ測定機械であってもよい。

図面の簡単な説明

図1は、遺伝子発現解析システムのブロック図である。

図2は、解析ソフトのフローチャートである。

図3は、交差検証成績CVの計算のフローチャートである。

図4は、変数選択の第1モデル構築手法のフローチャートである。

図5は、変数選択の第2モデル構築手法のフローチャートである。

図6は、変数選択の第3モデル構築手法のフローチャートである。

図7は、変数選択の第4モデル構築手法のフローチャートである。

図8は、変数選択の第5モデル構築手法のフローチャートである。

図 9 は、最小自乗法モデルの成績を示すグラフである。

図 10 は、DLBCL 患者の生存時間と診断指標のプロット各種比較の図である。

図 11 は、実施例 2 の DLBCL 患者の生存時間診断指標のプロットの図である。

図 12 は、実施例 3 の乳癌患者の生存時間診断指標のプロットの図である。

5 図 13 は、実施例 3 の乳癌患者の変数削除基準として $P \geq 0.0005$ を採用したときの生存時間診断指標のプロットの図である。

図 14 は、実施例 7 の乳癌患者の再発時間診断指標のプロットの図である。

図 15 は、実施例 7 の乳癌患者の変数削除基準として $P \geq 0.025$ を採用したときの再発時間診断指標のプロットの図である。

10 図 16 は、実施例 9 の遺伝的アルゴリズムによる部分最小自乗法モデルの最適化の様子を示す図である。

図 17 は、実施例 10 の階層型人工ニューラルネットワークにおける 4 つのトポロジーを示す図である。

15 図 18 は、実施例 11 の潜在変数を用いた比例ハザードモデルの乳癌患者の生存時間診断指標のグラフである。

図 19 は、実施例 11 の潜在変数を用いた比例ハザードモデルの乳癌患者の生存時間診断指標の予測値と計算値のグラフである。

発明を実施するための最良の形態

20 以下、添付の図面を参照して本発明の実施の形態を説明する。

以下に、選択された生体の状態と遺伝子発現の量および／または細胞内物質の量との相関モデルの決定について説明する。ここで、遺伝子発現の用語は、mRNA 発現(トランスクリプトーム)や、mRNA による翻訳の結果として生じる蛋白質(プロテオーム)を含むものとして用いる。また、細胞内物質の量とはここではたとえば、代謝中間体を含めた代謝産物全部であるメタボロームを意味する。たとえば、トランスクリプトーム(mRNA)やプロテオーム(蛋白質)の解析において、各サンプルデータは、生体の状態と遺伝子発現の量などからなる。各サンプルはたとえば 1000 個以上の膨大な遺伝子発現の量を含む。生体の状態は、たとえば病気のタイプまたは病気の診断指標であるが、より一般的には生体情報

25

であればよい。「病気の診断指標」には、病気の進行度合いのほか、病気のタイプ、重篤度、深刻度などの表現で表わされるものも含む。ここで、遺伝子発現の量などの測定データは膨大な情報量からなるので、コンピュータを用いた効率的な多変量解析が必要である。

- 5 データ収集において、予めいくつかのサンプルについて生体の状態（たとえば診断指標）を判定し、また、そのサンプルされたものから細胞液を獲得し、その細胞液中の多くの遺伝子産物の発現の量などを測定する。本発明の実施の形態のデータ解析では、こうして得られた遺伝子産物の発現の量などと生体の状態（たとえば診断指標）を入力し、相関モデル（たとえば部分最小自乗法モデル）を得る。ここで、コンピュータによる多変量解析プログラムを用いて、診断指標を目的変数とし、遺伝子発現の量および／または細胞内物質の量を説明変数とする因果関係型の解析を行なって、各説明変数の重要性や影響度に関する情報を得る。また、前記目的変数は、必ずしも測定値そのものである必要はなく、ロジット変換を行なった値や群を表す離散値を用いても良く、その場合、より有意な解析結果を得ることもできる。
- 10
- 15

- 本発明者らは、遺伝子発現による医療診断という分野において、データ解析における交差検証（cross validation）の成績を少なくとも独立変数のひとつとして持つ関数を最適化するように変数を選択することによって良好な相関モデル（たとえば部分最小自乗法モデル）が得られることを見出した。交差検証法では、手持ちのデータを複数群に分割し、その一部のデータ群（訓練集合）だけを使ってフィットしたモデルを用いて残る別のデータ群（テスト集合）を予測することによって、モデルの予測力を試す。通常の部分最小自乗法（PLS）においては潜在変数の次元選択に交差検証法が用いられているが、ここでは、部分最小自乗法において、潜在変数を1次元に固定し、1以上の入力変数（説明変数）を逐次取捨選択しながら、交差検証成績（たとえば平方和の予測誤差）を少なくとも独立変数のひとつとして持つ関数を最適化した。ただし本発明の効果は潜在変数の次元を1に限定するものではない。その結果、全変数を採用した場合には有意な相関モデルを得られなかった場合にも、良好でかつ予測力のある相関モデルが得られることが判明したのである。この交差検証法を用いた変数選択の逐次取捨選択
- 20
- 25

により、安定な相関モデルが得られる。また本発明者らは、関数形を適切に設定することによって説明変数を絞り込むことにより、部分最小自乗法以外の統計学又は多変量解析の良好な相関モデルを得ることが可能となり、それぞれ生体の状態を記述する目的変数にふさわしい相関モデルを得ることができることを見出した。なお、ここでいう「最適化」とは、交差検証成績が、説明変数を取捨選択するための、そのときの解析条件の範囲で、改善がみられなくなるまで改良したことを意味しており、交差検証成績がすべての説明変数の組合せの中で最適なものを見出したという意味ではない。この変数選択手法を用いると、病状を決定する因子を少数に特定し、廉価な診断用材料（DNAチップ、抗体チップ、DNA含有ベクターなど）を設計でき、それ自体独自の価値を持つものである。また、この変数選択手法は、予め設定される各種の変数選択条件と共に運用することが可能である。

上に述べたように、説明変数は、交差検証成績を基準に逐次取捨選択される。ここで、取捨選択のため、交差検証成績を少なくとも独立変数のひとつとして持つ関数を用いる。説明変数を追加する場合は、その説明変数について、前記関数が改善されなかったと判定された場合には当該説明変数を除外し、改善されたと判定された場合には当該説明変数を追加する。また、説明変数を除外する場合は、その説明変数について、前記関数が改善されなかったと判定された場合には当該説明変数を除外せず、改善されたと判定された場合には当該説明変数を除外する。ここで、1以上の説明変数を選択した場合に、交差検証成績評価は次のように進める。n個のサンプルからいくつかのサンプルを逐次除外して部分最小自乗法モデルを求め、各モデルにおいて除外したサンプルの遺伝子発現の量から予測される生体の状態を示す目的変数と、除外したサンプルの生体の状態を示す目的変数との各々の誤差の代表値を求める。「代表値」とは、和、平均、最大値、中位値、最頻値などのデータを特徴づける値をいう。そして、当該誤差の代表値を少なくともひとつの独立変数とする関数が小さくなった場合に、交差検証成績が改善されたと判定し、当該説明変数を追加または削除する。この交差検証成績評価を、説明変数を取捨選択しながら逐次繰り返して、前記関数を改善し続ける。改善されなくなれば交差検証成績を最適化したとして説明変数の取捨選択を終了する。

その結果、取捨選択により絞り込んだ数の説明変数からなる最適な部分最小自乗法モデルが得られる。具体的には、計算手段において計算される交差検証成績の数値指標として予想残差自乗和 (PRESS) を採用し、評価判定手段において予想残差自乗和の値が説明変数あたり一定の閾値以下の比率で小さくなる場合に、その説明変数を採用すると判定することにより、上記の処理は実行可能である。

因果関係型の解析手法においてはオーバーフィット (over fitting) を避けるための工夫が必要となる。ここでいうオーバーフィットとは、説明変数が多すぎるためにたまたま予測結果と実績とが一致するものの、本当の相関関係をとらえ損なっているため、モデルフィットに用いたデータ以外に予測能力を持たないことをいう。ここでは、相関モデルとして部分最小自乗法を用いるが、部分最小自乗法は次元圧縮とモデルフィットを同時に行なう強力な多変量解析手法であり、オーバーフィットの問題に比較的強いとされている。しかし遺伝子発現状態解析のように膨大な変数を扱う場合には、有意な結果が得られない事態に直面する。従来技術として説明したAlaiyaやKhanの手法は全変数モデルが有意に成立することを前提としているので、変数の絞り込みには一般的には適用できない。これに対し、本発明では、交差検証予測結果を最適にするように変数を絞り込むことにより、オーバーフィットを減らすことができた。また、本発明は、前記Khanの手法とは異なり、主成分分析などの前処理を介さない方法である。従来技術では、説明変数が膨大な場合には、有意なモデルを得ることができないことから、予め、全説明変数を基にたとえば、主成分分析などで次元圧縮する前処理をし、これによって得られた説明変数によって解析する方法が用いられる。しかし、この方法では、構成したモデルで予測を行なうためには、モデル構成の基となった全説明変数が必ず必要となり、たとえば、説明変数が遺伝子発現の量であれば、診断用遺伝子チップに担持する遺伝子としては、モデル構成に用いた遺伝子の全てが必要となるか、または別の手法を用いて変数選択することが必要となる。一方、本発明においては、説明変数の選択によって説明変数を絞り込んでいるので、たとえば、説明変数が遺伝子発現の量であれば、診断用遺伝子チップに担持する遺伝子は、選択された説明変数に相当する遺伝子を担持すれば良いことになる。

なお、Todeschiniらは、有機化合物の大気中の分解を予測するため、遺伝的ア

ルゴリズムによって交差検証成績を最適化するように変数選択を行ない、重回帰モデルを得ている (P. Gramatics, V. Consonni & R. Todeschini, Chemosphere 38(5), 1371-78 (1999))。53化合物と175記述子でモデル構築を行ない ($Q^2 = 0.79$)、7変数が選択され、98化合物の予測を行なった ($Q^2 = 0.75$)。交

5 差検証成績を最適化するように変数選択を行なっている点では、本実施形態と同様の手法である。しかし、重回帰モデルを採用しているために、説明変数の選択過程を通じて選択される変数は少数個にとどまらざるを得ず、複数の遺伝子発現の量および／または細胞内物質の量の解析には適用できない。本発明者らの調査した範囲では、 Q^2 やPRESS値を最適化する方法では、選抜される説明変数は百程度

10 から数百程度にわたり、重回帰モデルでは解析が不能となる。またTodeschiniらは、説明変数を絞り込むための有効な方法について言及していない。これは、もともとの説明変数の候補がたかだか175個であり、説明変数を絞り込むために特別の工夫をする必要がないからである。遺伝子発現解析の分野はこれとは全く異なり、数十から数百のサンプル数に対して、数百から数千、数万の説明変数候補が存在する。したがってこれまでとは異なる工夫が必要となる。

15

本実施形態では、生体の状態と複数の遺伝子発現の量および／または細胞内物質の量との相関モデルを決定するとき、交差検証成績を少なくとも独立変数のひとつとして持つ関数を最適化させるように説明変数を逐次追加・除外することによって、説明変数を選抜して、良好な相関モデルを得る。このようなアプローチ

20 の優位性は、下記の実施例から推測されるように、次のとおりである。

- 1) 病気や生体現象の背後で働いている重要な遺伝子やメカニズムを推定／特定でき、理解が深まる。
- 2) 重要な遺伝子産物や細胞内物質だけに絞った廉価な診断用材料 (DNAチップ、抗体チップなど) の設計が可能になる。

25 本実施形態では、交差検証成績を少なくとも独立変数のひとつとして持つ関数を最適化するように説明変数を段階的に取捨選択するが、たとえば具体的には、ステップワイズ(step wise)法に代表される説明変数を選択する選択手段と、リーブ・ワン・アウト(leave-one-out)法に代表される交差検証法に部分最小自乗法を適用して計算する計算手段と、前記計算手段の結果を評価し、説明変数の採

用、不採用を判定する評価判定手段とを組合せて用いる。すなわち、 m 個の説明変数の中から1以上の説明変数を選択し、次いで、部分最小自乗法を実行して交差検証成績を計算し、さらに、該計算結果を評価して、選択した説明変数の採用、不採用を判定する。この評価判定では、計算手段の結果から、各計算において除外したサンプルの遺伝子発現から予測される生体の状態を示す目的変数値と、前記除外したサンプルの生体の状態を示す目的変数値との誤差の代表値を求め、少なくとも当該誤差の代表値を独立変数として持つ関数である当該誤差の代表値の単調減少関数の値が小さくなった場合に説明変数の取捨選択を判定する。このように、選択手段と計算手段と評価判定手段とを用いて、少なくとも部分最小自乗法モデルの交差検証成績を独立変数として持つ関数を改善し続けて、その改善がみられなくなるまで改良し、部分最小自乗法モデルを決定する。なお、本実施形態では、サンプルを1個ずつ逐次除外している(リーブ・ワン・アウト法)が、その代わりに、複数のサンプルを除外して交差検証成績を評価してもよい(リーブ・ n ・アウト法)し、また、Khan et al.により用いられた3分割法(three-fold)等の他の方法を用いることもできる。3分割法では、説明変数をランダムにシャッフルして3つのグループに分ける。その中の2つのグループを用いてモデルを構成し、残りの1つのグループでモデルを評価する。また、説明変数の選択方法としてはステップワイズ法、非線形アルゴリズム(たとえば遺伝的アルゴリズムなど)を用いてもよく、変数選択に関して予め何らかの条件が分っていれば、それに応じて探索範囲を限定できる。

次に、データの収集と解析について具体的に説明する。図1は、遺伝子発現解析システムを示す。データ収集のため、予めいくつかのサンプルについて診断指標(たとえば病気のタイプないし進行度合いを含む)を判定し、また、そのサンプルされたものから細胞液を獲得し、DNAチップを用いてその細胞液中の多くの遺伝子産物の発現の量を測定する。測定には、共焦点型レーザスキャナ(たとえばAffymetrix社、428アレイスキャナ)10を用いる。吸光度によりmRNAの量が測定される。このデータ収集は公知の方法である。測定データは、コンピュータ12に送られ解析される。コンピュータ12は、CPU14を備えた通常の構成のコンピュータであり、それに接続される記憶装置(たとえばハードデ

ディスク装置) 16の記録媒体(たとえばハードディスク)には、測定データ18
や解析ソフト20が格納される。この解析ソフト20を用いてデータ18が解析
され、生体の状態と遺伝子発現の量などとの相関モデルが決定される。

5 なお、説明変数の選択と、交差検証法に部分最小自乗法を適用する計算とを複
数のコンピュータで実行させてもよい。交差検証予測の計算を複数個のコンピ
ュータに分散させることで計算を加速することができる。

図2は、コンピュータ12により実行される、生体の状態と遺伝子発現の量な
どとの相関モデルを得るためのデータ解析ソフト20のフローチャートを示す。
ここでは簡単に説明するため、少なくとも部分最小自乗法モデルの交差検証成績
10 を独立変数として持つ関数としてPRESSを採用しているが、発明の範囲を限定す
るものでなく、実施例2~5においては別の関数を採用している。まず、相関モ
デル作成用のデータを入力する(S10)。データはたとえばDNAチップを用
いて収集したものである。入力データ(サンプル集合)は、それぞれ目的変数

15 (たとえば診断指標)とm個(たとえば2000個)の説明変数(たとえば遺伝
子発現の量)からなる。また、場合によっては、上述のデータ(訓練集合)以外に、
テスト集合のデータを入力する。ここでテスト集合とは交差検証の評価のための
データ群を意味するのではなく、モデル決定が終了した後にモデルの予測力をテ
ストするためのデータ群である。

20 まず、初期設定として、選択された説明変数の数を0とし、交差検証成績CVの
最良値CV₀を $-\infty$ とする(S12)。次に、説明変数の選択を行う。まず、説明
変数を指す番号iを1とし(S14)、第i変数(遺伝子発現の量)を仮に採用
して(S16)、部分最小自乗法を実行し、交差検証成績CVを計算する(S18、
図3参照)。ここで、リーブ・ワン・アウト処理を用いる。これは、たとえば5
0個のサンプルからなる訓練集合において、1番から50番の全てを順次1個づ
25 つ除いて残りの49個のサンプルで予測した結果と、その時除いた1個の結果と
を比較し、その誤差が大きい場合に、仮に選択した説明変数(第i変数)が適して
いないと判断する手法である。もし、得られた成績CVが現在の最良値CV₀より最
適化されれば(S20でYES)、第i変数を採用し、かつ、成績CVを新しい
最良値CV₀に更新する(S22)。しかし、得られた成績CVが最良値CV₀より大

きくなければ (S 2 0 で NO)、第 i 変数を採用しない (S 2 4)。そして、ステップ S 1 4 に戻り、同様の処理を繰り返す。この処理を交差検証成績 CV が改善されなくなる (S 2 6 で NO) まで繰り返す。ここで、相関モデルに採用する説明変数については 1 つずつ段階的に増加 (追加) または減少 (除外) して成績 CV を評価判定している。すなわち、全体としての合致度合いがよくなるように各説明変数を解析に加えるかどうかを逐次判定しながら、説明変数の取捨選択を行い、これを、全体としての合致度合いがよくなるまで繰り返す。以上の処理で改善があると、ふたたびステップ S 1 4 の初め ($i=1$) に戻り、それまでに選択されている説明変数を基に、さらに説明変数の選択を繰り返す。なお、ここではモデルの予測力を判断するために、訓練集合とテスト集合とに予め分割しておいたデータ集合を用いてデータ解析しており、上述の解析は、訓練集合を用いて行なった結果であるので、この結果からテスト集合について予測を行い、実測データとの合致度を評価 (S 2 8) している。このような評価は必ずしも必要でないが、予測力を判断するには有効である。

図 3 は、リーブ・ワン・アウト処理を含む交差検証成績 CV の計算 (図 2、S 1 8) のフローチャートを示す。ここで、選択された変数について交差検証成績が計算される。まず、PRESS の初期値を 0 とする (S 1 8 0)。次に、 n 個の集合内のサンプルを指す番号 j を 1 とし (S 1 8 2)、第 j サンプル以外の $n-1$ 個のサンプルで部分最小自乗法を実行し (S 1 8 4)、第 j サンプルの目的変数を予測する (S 1 8 6)。差の自乗を計算して PRESS に加算する (S 1 9 0)。次に番号 j を 1 増加し (S 1 8 2)、同様の処理をおこなう。これを番号 $j=n$ まで各サンプルについて繰り返す。得られた PRESS は、1 個のサンプルを順次除外して計算した予測値と実測値との差の平方和であり、予測誤差を表わす量である。この予測残差自乗和 PRESS の符号を変えたものを交差検証成績 CV とする (S 1 9 2)。

本実施形態では、交差検証法を用いて、入力変数 (説明変数) を段階的に 1 つずつ追加・除外しながら、交差検証成績 ($CV=-PRESS$) を最適化する。ここで、説明変数の段階的な追加・除外の内容を理解しやすくするため、以下で、さらに具体的に 5 つのモデル構築手法について説明する。これらは、説明変数の逐次的

な選択の手順が異なる。

図4は、第1のモデル構築手法を示す。データ集合においてどの説明変数も選択されていない状態を初期状態とする(S112)。次に、1番目の説明変数から最後(m番目)の説明変数までの未だ選択されていない説明変数ごとに逐次、その説明変数を選択した場合に交差検証成績が改善するかどうかを、リーブ・ワ
ン・アウト処理を用いた交差検証成績評価ステップ(S118)を繰り返しながら判定(S120)し、改善する場合にはその説明変数を追加する(S114~S124)。そのような改善と追加がなくなる(S126でNO)まで、1番目の説明変数から上記逐次判定操作を繰り返す。

さらに詳しく説明すると、まず、初期設定として、選択された説明変数の数N
Pを0とし、交差検証成績CVの最良値 CV_0 を $-\infty$ とする(S112)。次に、説明変数の選択を行う。まず、変数iを1とし(S114)、第i変数を仮に採用する(S116)。ただし、第i変数がすでに採用されていれば(S115でYES)、ステップS114に戻る。次に、部分最小自乗法を実行し、交差検証成績CVを計算する(S118)。ここで、リーブ・ワン・アウト処理を用いる。もし、得られた成績CVが現在の最良値 CV_0 より最適化されれば(S120でYES)、第i変数を採用し、かつ、成績CVを新しい最良値 CV_0 に更新する(S122)。しかし、得られた成績CVが最良値 CV_0 より大きくなければ(S120でNO)、第i変数を採用しない(S124)。そして、ステップS114に戻り、同様の処理を繰り返す。この処理を交差検証成績CVが改善されなくなる(S126でNO)まで繰り返す。以上の処理で改善があると、ふたたびステップS114に戻り、新しいループを開始する。ここで、それまでに選択されている変数を基に、さらに変数の選択を繰り返す。こうして、データ集合を用いて選択された変数を用いた相関モデルが得られる。

図5は、第2のモデル構築手法を示す。この手法では、全ての説明変数が選択されている状態を初期状態とする(S212)。次に、1番目の説明変数から最後(m番目)の説明変数までの選択されている説明変数ごとに逐次、その説明変数を除外した場合に交差検証成績が改善するかどうかを、リーブ・ワン・アウト処理を用いた交差検証成績評価ステップ(S218)を繰り返しながら判定(S2

20)し、改善する場合にはその説明変数を除外する(S 2 1 4～S 2 2 4)。そのような改善と除外がなくなる(S 2 2 6でNO)まで、1番目の説明変数から上記逐次判定操作を繰り返す。

さらに詳しく説明すると、まず、初期設定として、選択された説明変数の数N
5 Pをmとし、交差検証成績CVの最良値 CV_0 を $-\infty$ とする(S 2 1 2)。すなわち、すべての説明変数を選択する。次に、説明変数の選択を行う。まず、変数iを1とし(S 2 1 4)、第i変数を仮に除外する(S 2 1 6)。ただし、第i変数がすでに除外されていれば(S 2 1 5でYES)、ステップS 2 1 4に戻る。部分
10 最小自乗法を実行し、交差検証成績CVを計算する(S 2 1 8)。ここで、リーブ・ワン・アウト処理を用いる。もし、得られた成績CVが現在の最良値 CV_0 より最適化されれば(S 2 2 0でYES)、第i変数を除外し、かつ、成績CVを新しい最良値 CV_0 に更新する(S 2 2 2)。しかし、得られた成績CVが最良値 CV_0 より大きくなければ(S 2 2 0でNO)、第i変数を除外しない(S 2 2 4)。そして、ステップS 2 1 4に戻り、同様の処理を繰り返す。この処理を交差検証成績CVが改善されなくなる(S 2 2 6でNO)まで繰り返す。以上の処理で改善
15 があると、ふたたびステップS 2 1 4に戻り、新しいループを開始する。ここで、それまでに選択されている変数を基に、さらに変数の選択を繰り返す。こうして、データ集合を用いて選択された変数を用いた相関モデルが得られる。

図6は、第3のモデル構成手法を示す。この手法は、第1と第2の手法の直列的な組合せである。まず、どの説明変数も選択されていない状態を初期状態とする(S 1 1 2)。次に、1番目の説明変数から最後(m番目)の説明変数までの
20 未だ選択されていない説明変数ごとに逐次、その説明変数を選択した場合に交差検証成績が改善するかどうかを、リーブ・ワン・アウト処理を用いた交差検証成績評価ステップを繰り返しながら判定し、改善する場合にはその説明変数を追加
25 選択し、そのような改善と追加がなくなるまで1番目の説明変数から上記逐次判定操作を繰り返す(S 1 1 4～S 1 2 6)。次に、1番目の説明変数から最後

(m番目)の説明変数までの選択されている説明変数ごとに逐次、その説明変数を除外した場合に交差検証成績が改善するかどうかを、リーブ・ワン・アウト処理を用いた交差検証成績評価ステップを繰り返しながら判定し、改善する場合に

はその説明変数を除外し、そのような改善と除外がなくなるまで、1番目の説明変数から上記逐次判定操作を繰り返す（S 2 1 4～S 2 2 6）。

図 7 は、第 4 のモデル構築手法を示す。この手法は、第 3 の手法の変形である。まず、どの説明変数も選択されていない状態を初期状態とする（S 1 1 2）。次に、1番目の説明変数から最後（m 番目）の説明変数までの未だ選択されていない説明変数ごとに逐次、その説明変数を選択した場合に交差検証成績が改善するかどうかを、リーブ・ワン・アウト処理を用いた交差検証成績評価ステップ（S 1 1 8）を繰り返しながら判定（S 1 2 0）し、改善する場合にはその説明変数を追加選択する（S 1 1 4～S 1 2 4）。そのような改善と追加がなくなる（S 1 2 6 で NO）まで、1番目の説明変数から上記逐次判定操作を繰り返す。次に、1番目の説明変数から最後（m 番目）の説明変数までの選択されている説明変数ごとに逐次、その説明変数を除外した場合に交差検証成績が改善するかどうかを、リーブ・ワン・アウト処理を用いた交差検証成績評価ステップ（S 2 1 8）を繰り返しながら判定（S 2 2 0）し、改善する場合にはその説明変数を除外する（S 2 1 4～2 2 4）。そのような改善と除外がなくなる（S 2 2 6 で NO）まで、1番目の説明変数から上記逐次判定操作を繰り返す。上記逐次判定追加改善ステップまたは上記逐次判定除外改善ステップで少なくとも一度改善があれば（S 2 2 7 で YES）、ステップ S 1 1 2 に戻り、上記操作（S 1 1 2～S 2 2 7）を繰り返す。これを改善がなくなる（S 2 2 7 で NO）までおこなう。

図 8 は、第 5 のモデル構築手法を示す。この手法は、第 1 と第 2 のスキームの並列的な組合せである。どの説明変数も選択されていない状態を初期状態とする（S 1 1 2）。次に、1番目の説明変数から最後（m 番目）の説明変数までの説明変数ごとに逐次、その説明変数が選択されていない場合にはその説明変数を選択した場合に交差検証成績が改善するかどうかを、リーブ・ワン・アウト処理を用いた交差検証成績評価ステップ（S 1 1 8）を繰り返しながら判定（S 1 2 0）し、改善する場合にはその説明変数を追加する（S 1 1 4～S 1 2 4）。また、選択する説明変数ごとに、その説明変数がすでに選択されている場合には、その説明変数を除外した場合に交差検証成績が改善するかどうかを、リーブ・ワン・アウト処理を用いた交差検証成績評価ステップ（S 2 1 8）を繰り返しながら

ら判定 (S 2 2 0) し、改善する場合にはその説明変数を除外する (S 2 1 6 ~ S 2 2 4)。そのような改善と追加または除外がなくなる (S 1 2 6 で NO) ま
で、1 番目の説明変数から上記逐次判定操作を繰り返す。

次に、第 4 のモデル構築手法 (図 7) を適用した場合を、表 1 のデータ集合を例
として説明する。このデータ集合に対して、部分最小自乗法による解析を用いて
5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100 105 110 115 120 125 130 135 140 145 150 155 160 165 170 175 180 185 190 195 200 205 210 215 220 225 230 235 240 245 250 255 260 265 270 275 280 285 290 295 300 305 310 315 320 325 330 335 340 345 350 355 360 365 370 375 380 385 390 395 400 405 410 415 420 425 430 435 440 445 450 455 460 465 470 475 480 485 490 495 500 505 510 515 520 525 530 535 540 545 550 555 560 565 570 575 580 585 590 595 600 605 610 615 620 625 630 635 640 645 650 655 660 665 670 675 680 685 690 695 700 705 710 715 720 725 730 735 740 745 750 755 760 765 770 775 780 785 790 795 800 805 810 815 820 825 830 835 840 845 850 855 860 865 870 875 880 885 890 895 900 905 910 915 920 925 930 935 940 945 950 955 960 965 970 975 980 985 990 995 1000 1005 1010 1015 1020 1025 1030 1035 1040 1045 1050 1055 1060 1065 1070 1075 1080 1085 1090 1095 1100 1105 1110 1115 1120 1125 1130 1135 1140 1145 1150 1155 1160 1165 1170 1175 1180 1185 1190 1195 1200 1205 1210 1215 1220 1225 1230 1235 1240 1245 1250 1255 1260 1265 1270 1275 1280 1285 1290 1295 1300 1305 1310 1315 1320 1325 1330 1335 1340 1345 1350 1355 1360 1365 1370 1375 1380 1385 1390 1395 1400 1405 1410 1415 1420 1425 1430 1435 1440 1445 1450 1455 1460 1465 1470 1475 1480 1485 1490 1495 1500 1505 1510 1515 1520 1525 1530 1535 1540 1545 1550 1555 1560 1565 1570 1575 1580 1585 1590 1595 1600 1605 1610 1615 1620 1625 1630 1635 1640 1645 1650 1655 1660 1665 1670 1675 1680 1685 1690 1695 1700 1705 1710 1715 1720 1725 1730 1735 1740 1745 1750 1755 1760 1765 1770 1775 1780 1785 1790 1795 1800 1805 1810 1815 1820 1825 1830 1835 1840 1845 1850 1855 1860 1865 1870 1875 1880 1885 1890 1895 1900 1905 1910 1915 1920 1925 1930 1935 1940 1945 1950 1955 1960 1965 1970 1975 1980 1985 1990 1995 2000 2005 2010 2015 2020 2025 2030 2035 2040 2045 2050 2055 2060 2065 2070 2075 2080 2085 2090 2095 2100 2105 2110 2115 2120 2125 2130 2135 2140 2145 2150 2155 2160 2165 2170 2175 2180 2185 2190 2195 2200 2205 2210 2215 2220 2225 2230 2235 2240 2245 2250 2255 2260 2265 2270 2275 2280 2285 2290 2295 2300 2305 2310 2315 2320 2325 2330 2335 2340 2345 2350 2355 2360 2365 2370 2375 2380 2385 2390 2395 2400 2405 2410 2415 2420 2425 2430 2435 2440 2445 2450 2455 2460 2465 2470 2475 2480 2485 2490 2495 2500 2505 2510 2515 2520 2525 2530 2535 2540 2545 2550 2555 2560 2565 2570 2575 2580 2585 2590 2595 2600 2605 2610 2615 2620 2625 2630 2635 2640 2645 2650 2655 2660 2665 2670 2675 2680 2685 2690 2695 2700 2705 2710 2715 2720 2725 2730 2735 2740 2745 2750 2755 2760 2765 2770 2775 2780 2785 2790 2795 2800 2805 2810 2815 2820 2825 2830 2835 2840 2845 2850 2855 2860 2865 2870 2875 2880 2885 2890 2895 2900 2905 2910 2915 2920 2925 2930 2935 2940 2945 2950 2955 2960 2965 2970 2975 2980 2985 2990 2995 3000 3005 3010 3015 3020 3025 3030 3035 3040 3045 3050 3055 3060 3065 3070 3075 3080 3085 3090 3095 3100 3105 3110 3115 3120 3125 3130 3135 3140 3145 3150 3155 3160 3165 3170 3175 3180 3185 3190 3195 3200 3205 3210 3215 3220 3225 3230 3235 3240 3245 3250 3255 3260 3265 3270 3275 3280 3285 3290 3295 3300 3305 3310 3315 3320 3325 3330 3335 3340 3345 3350 3355 3360 3365 3370 3375 3380 3385 3390 3395 3400 3405 3410 3415 3420 3425 3430 3435 3440 3445 3450 3455 3460 3465 3470 3475 3480 3485 3490 3495 3500 3505 3510 3515 3520 3525 3530 3535 3540 3545 3550 3555 3560 3565 3570 3575 3580 3585 3590 3595 3600 3605 3610 3615 3620 3625 3630 3635 3640 3645 3650 3655 3660 3665 3670 3675 3680 3685 3690 3695 3700 3705 3710 3715 3720 3725 3730 3735 3740 3745 3750 3755 3760 3765 3770 3775 3780 3785 3790 3795 3800 3805 3810 3815 3820 3825 3830 3835 3840 3845 3850 3855 3860 3865 3870 3875 3880 3885 3890 3895 3900 3905 3910 3915 3920 3925 3930 3935 3940 3945 3950 3955 3960 3965 3970 3975 3980 3985 3990 3995 4000 4005 4010 4015 4020 4025 4030 4035 4040 4045 4050 4055 4060 4065 4070 4075 4080 4085 4090 4095 4100 4105 4110 4115 4120 4125 4130 4135 4140 4145 4150 4155 4160 4165 4170 4175 4180 4185 4190 4195 4200 4205 4210 4215 4220 4225 4230 4235 4240 4245 4250 4255 4260 4265 4270 4275 4280 4285 4290 4295 4300 4305 4310 4315 4320 4325 4330 4335 4340 4345 4350 4355 4360 4365 4370 4375 4380 4385 4390 4395 4400 4405 4410 4415 4420 4425 4430 4435 4440 4445 4450 4455 4460 4465 4470 4475 4480 4485 4490 4495 4500 4505 4510 4515 4520 4525 4530 4535 4540 4545 4550 4555 4560 4565 4570 4575 4580 4585 4590 4595 4600 4605 4610 4615 4620 4625 4630 4635 4640 4645 4650 4655 4660 4665 4670 4675 4680 4685 4690 4695 4700 4705 4710 4715 4720 4725 4730 4735 4740 4745 4750 4755 4760 4765 4770 4775 4780 4785 4790 4795 4800 4805 4810 4815 4820 4825 4830 4835 4840 4845 4850 4855 4860 4865 4870 4875 4880 4885 4890 4895 4900 4905 4910 4915 4920 4925 4930 4935 4940 4945 4950 4955 4960 4965 4970 4975 4980 4985 4990 4995 5000 5005 5010 5015 5020 5025 5030 5035 5040 5045 5050 5055 5060 5065 5070 5075 5080 5085 5090 5095 5100 5105 5110 5115 5120 5125 5130 5135 5140 5145 5150 5155 5160 5165 5170 5175 5180 5185 5190 5195 5200 5205 5210 5215 5220 5225 5230 5235 5240 5245 5250 5255 5260 5265 5270 5275 5280 5285 5290 5295 5300 5305 5310 5315 5320 5325 5330 5335 5340 5345 5350 5355 5360 5365 5370 5375 5380 5385 5390 5395 5400 5405 5410 5415 5420 5425 5430 5435 5440 5445 5450 5455 5460 5465 5470 5475 5480 5485 5490 5495 5500 5505 5510 5515 5520 5525 5530 5535 5540 5545 5550 5555 5560 5565 5570 5575 5580 5585 5590 5595 5600 5605 5610 5615 5620 5625 5630 5635 5640 5645 5650 5655 5660 5665 5670 5675 5680 5685 5690 5695 5700 5705 5710 5715 5720 5725 5730 5735 5740 5745 5750 5755 5760 5765 5770 5775 5780 5785 5790 5795 5800 5805 5810 5815 5820 5825 5830 5835 5840 5845 5850 5855 5860 5865 5870 5875 5880 5885 5890 5895 5900 5905 5910 5915 5920 5925 5930 5935 5940 5945 5950 5955 5960 5965 5970 5975 5980 5985 5990 5995 6000 6005 6010 6015 6020 6025 6030 6035 6040 6045 6050 6055 6060 6065 6070 6075 6080 6085 6090 6095 6100 6105 6110 6115 6120 6125 6130 6135 6140 6145 6150 6155 6160 6165 6170 6175 6180 6185 6190 6195 6200 6205 6210 6215 6220 6225 6230 6235 6240 6245 6250 6255 6260 6265 6270 6275 6280 6285 6290 6295 6300 6305 6310 6315 6320 6325 6330 6335 6340 6345 6350 6355 6360 6365 6370 6375 6380 6385 6390 6395 6400 6405 6410 6415 6420 6425 6430 6435 6440 6445 6450 6455 6460 6465 6470 6475 6480 6485 6490 6495 6500 6505 6510 6515 6520 6525 6530 6535 6540 6545 6550 6555 6560 6565 6570 6575 6580 6585 6590 6595 6600 6605 6610 6615 6620 6625 6630 6635 6640 6645 6650 6655 6660 6665 6670 6675 6680 6685 6690 6695 6700 6705 6710 6715 6720 6725 6730 6735 6740 6745 6750 6755 6760 6765 6770 6775 6780 6785 6790 6795 6800 6805 6810 6815 6820 6825 6830 6835 6840 6845 6850 6855 6860 6865 6870 6875 6880 6885 6890 6895 6900 6905 6910 6915 6920 6925 6930 6935 6940 6945 6950 6955 6960 6965 6970 6975 6980 6985 6990 6995 7000 7005 7010 7015 7020 7025 7030 7035 7040 7045 7050 7055 7060 7065 7070 7075 7080 7085 7090 7095 7100 7105 7110 7115 7120 7125 7130 7135 7140 7145 7150 7155 7160 7165 7170 7175 7180 7185 7190 7195 7200 7205 7210 7215 7220 7225 7230 7235 7240 7245 7250 7255 7260 7265 7270 7275 7280 7285 7290 7295 7300 7305 7310 7315 7320 7325 7330 7335 7340 7345 7350 7355 7360 7365 7370 7375 7380 7385 7390 7395 7400 7405 7410 7415 7420 7425 7430 7435 7440 7445 7450 7455 7460 7465 7470 7475 7480 7485 7490 7495 7500 7505 7510 7515 7520 7525 7530 7535 7540 7545 7550 7555 7560 7565 7570 7575 7580 7585 7590 7595 7600 7605 7610 7615 7620 7625 7630 7635 7640 7645 7650 7655 7660 7665 7670 7675 7680 7685 7690 7695 7700 7705 7710 7715 7720 7725 7730 7735 7740 7745 7750 7755 7760 7765 7770 7775 7780 7785 7790 7795 7800 7805 7810 7815 7820 7825 7830 7835 7840 7845 7850 7855 7860 7865 7870 7875 7880 7885 7890 7895 7900 7905 7910 7915 7920 7925 7930 7935 7940 7945 7950 7955 7960 7965 7970 7975 7980 7985 7990 7995 8000 8005 8010 8015 8020 8025 8030 8035 8040 8045 8050 8055 8060 8065 8070 8075 8080 8085 8090 8095 8100 8105 8110 8115 8120 8125 8130 8135 8140 8145 8150 8155 8160 8165 8170 8175 8180 8185 8190 8195 8200 8205 8210 8215 8220 8225 8230 8235 8240 8245 8250 8255 8260 8265 8270 8275 8280 8285 8290 8295 8300 8305 8310 8315 8320 8325 8330 8335 8340 8345 8350 8355 8360 8365 8370 8375 8380 8385 8390 8395 8400 8405 8410 8415 8420 8425 8430 8435 8440 8445 8450 8455 8460 8465 8470 8475 8480 8485 8490 8495 8500 8505 8510 8515 8520 8525 8530 8535 8540 8545 8550 8555 8560 8565 8570 8575 8580 8585 8590 8595 8600 8605 8610 8615 8620 8625 8630 8635 8640 8645 8650 8655 8660 8665 8670 8675 8680 8685 8690 8695 8700 8705 8710 8715 8720 8725 8730 8735 8740 8745 8750 8755 8760 8765 8770 8775 8780 8785 8790 8795 8800 8805 8810 8815 8820 8825 8830 8835 8840 8845 8850 8855 8860 8865 8870 8875 8880 8885 8890 8895 8900 8905 8910 8915 8920 8925 8930 8935 8940 8945 8950 8955 8960 8965 8970 8975 8980 8985 8990 8995 9000 9005 9010 9015 9020 9025 9030 9035 9040 9045 9050 9055 9060 9065 9070 9075 9080 9085 9090 9095 9100 9105 9110 9115 9120 9125 9130 9135 9140 9145 9150 9155 9160 9165 9170 9175 9180 9185 9190 9195 9200 9205 9210 9215 9220 9225 9230 9235 9240 9245 9250 9255 9260 9265 9270 9275 9280 9285 9290 9295 9300 9305 9310 9315 9320 9325 9330 9335 9340 9345 9350 9355 9360 9365 9370 9375 9380 9385 9390 9395 9400 9405 9410 9415 9420 9425 9430 9435 9440 9445 9450 9455 9460 9465 9470 9475 9480 9485 9490 9495 9500 9505 9510 9515 9520 9525 9530 9535 9540 9545 9550 9555 9560 9565 9570 9575 9580 9585 9590 9595 9600 9605 9610 9615 9620 9625 9630 9635 9640 9645 9650 9655 9660 9665 9670 9675 9680 9685 9690 9695 9700 9705 9710 9715 9720 9725 9730 9735 9740 9745 9750 9755 9760 9765 9770 9775 9780 9785 9790 9795 9800 9805 9810 9815 9820 9825 9830 9835 9840 9845 9850 9855 9860 9865 9870 9875 9880 9885 9890 9895 9900 9905 9910 9915 9920 9925 9930 9935 9940 9945 9950 9955 9960 9965 9970 9975 9980 9985 9990 9995 10000 10005 10010 10015 10020 10025 10030 10035 10040 10045 10050 10055 10060 10065 10070 10075 10080 10085 10090 10095 10100 10105 10110 10115 10120 10125 10130 10135 10140 10145 10150 10155 10160 10165 10170 10175 10180 10185 10190 10195 10200 10205 10210 10215 10220 10225 10230 10235 10240 10245 10250 10255 10260 10265 10270 10275 10280 10285 10290 10295 10300 10305 10310 10315 10320 10325 10330 10335 10340 10345 10350 10355 10360 10365 10370 10375 10380 10385 10390 10395 10400 10405 10410 10415 10420 10425 10430 10435 10440 10445 10450 10455 10460 10465 10470 10475 10480 10485 10490 10495 10500 10505 10510 10515 10520 10525 10530 10535 10540 10545 10550 10555 10560 10565 10570 10575 10580 10585 10590 10595 10600 10605 10610 10615 10620 10625 10630 10635 10640 10645 10650 10655 10660 10665 10670 10675 10680 10685 10690 10695 10700 10705 10710 10715 10720 10725 10730 10735 10740 10745 10750 10755 10760 10765 10770 10775 10780 10785 10790 10795 10800 10805 10810 10815 10820 10825 10830 10835 10840 10845 10850 10855 10860 10865 10870 10875 10880 10885 10890 10895 10900 10905 10910 10915 10920 10925 10930 10935 10940 10945 10950 10955 10960 10965 10970 10975 10980 10985 10990 10995 11000 11005 11010 11015 11020 11025 11030 11035 11040 11045 11050 11055 11060 11065 11070 11075 11080 11085 11090 11095 11100 11105 11110 11115 11120 11125 11130 11135 11140 11145 11150 11155 11160 11165 11170 11175 11180 11185 11190 11195 11200 11205 11210 11215 11220 11225 11230 11235 11240 11245 11250 11255 11260 11265 11270 11275 11280 11285 11290 11295 11300 11305 11310 11315 11320 11325 11330 11335 11340 11345 11350 11355 11360 11365 11370 11375 11380 11385 11390 11395 11400 11405 11410 11415 11420 11425 11430 11435 11440 11445 11450 11455 11460 11465 11470 11475 11480 11485 11490 11495 11500 11505 11510 11515 11520 11525 11530 11535 11540 11545 11550 11555 11560 11565 11570 11575 11580 11585 11590 11595 11600 11605 11610 11615 11620 11625 11630 11635 11640 11645 11650 11655 11660 11665 11670 11675 11680 11685 11690 11695 11700 11705 11710 11715 11720 11725 11730 11735 11740 11745 11750 11755 11760 11765 11770 11775 11780 11785 11790 11795 11800 11805 11810 11815 11820 11825 11830 11835 11840 11845 11850 11855 11860 11865 11870 11875 11880 11885 11890 11895 11900 11905 11910 11

表2 表1のデータについての10の段階での変数選択結果

0		∞	-
1 追加	p20	0.111	p20
2 追加	p18	0.090	p18 & p20
3 追加	p16	0.073	p16 & p18 & p20
4 追加	p10	0.073	p10 & p16 & p18 & p20
5 追加	p6	0.062	p6 & p10 & p16 & p18 & p20
6 追加	p3	0.060	p3 & p6 & p10 & p16 & p18 & p20
7 追加	p12	0.055	p3 & p6 & p10 & p12 & p16 & p18 & p20
8 除外	p20	0.053	p3 & p6 & p10 & p12 & p16 &
9 除外	p10	0.050	p3 & p6 & p12 & p16 & p18
10 追加	p13	0.048	p3 & p6 & p12 & p13 & p16 & p15

先に述べたように、変数はp20からp2まで逆の順で処理する。表2は、表1のサンプルについて、左端の数字は、変数の取捨選択で改善がみられた10の段階を示す。なお、0は初期状態を意味する。次の列の「追加」と「除外」は、追加のループと除外のループの処理であることを意味する。次の列の変数は、追加または除外された変数を示す。次の列は、交差検証成績(PRESSをサンプル数で割ったもの)を示す。右端の列は、その段階で選択されている変数を示す。

初期状態では、変数は全くない状態であり、PRESSは ∞ である。表2に示すように、最初、p20を説明変数として採用すると、PRESS=0.111となり、初期値に比べて改善されるので、説明変数p20の追加を実施する。次に、変数p19を加えてp19とp20の2つを説明変数とすると、PRESS=0.129となり改善をもたらさないもので、p19は追加しない。次に、説明変数p18を加えるとPRESS=0.090となり、改善するので、p18を追加し、p18とp20を説明変数とする。以下同様に表2に示すように続く。(ここで、p10を追加採用するのは、小数点以下4桁目で改善されているためである。)説明変数p20～p2の1回目のループを終了した時点で、説明変数が

p3、p6、p10、p16、p18およびp20となり、PRESS=0.60となる。2回目のループでは、説明変数p12が追加され、PRESS=0.55となる。3回目のループでは追加による改善がなく、ひとまずS 1 1 4～S 1 2 6の追加処理を終了し、S 2 1 4に移る。この時点での部分最小自乗法のフィットならびにリーブ・ワン・アウト予測状況は表3のとおりである。

表3は、10のサンプルについて、表2の7で示す段階まで処理が進んだ時点での部分最小自乗法のフィットならびにリーブ・ワン・アウト予測状況を示す。ここで、モデル予測とリーブ・ワン・アウト予測のそれぞれにおいて、計算値と実測値との誤差を示す。さらに、その下側に、誤差の自乗平均、相関係数Rの自乗および予測相関係数Qの自乗を示す。

表3 表2の段階7での処理結果

モデル予測値				リーブワンアウト予測	
#	実測値	計算値	誤差	計算値	誤差
1	0.713	0.757	-0.044	0.693	0.020
2	0.133	-0.056	0.189	-0.051	0.184
3	0.545	0.497	0.048	0.480	0.065
4	0.752	0.646	0.106	0.495	0.257
5	0.900	0.687	0.214	0.557	0.343
6	0.455	0.489	-0.034	0.512	-0.057
7	0.427	0.624	-0.198	0.672	-0.245
8	0.042	0.349	-0.307	0.517	-0.475
9	0.935	0.865	0.070	0.782	0.153
10	0.154	0.197	-0.044	0.285	-0.132
0.093			0.024	0.055	
R ² =0.744			Q ² =0.407		

次に、ステップS 2 1 4から始まる除外処理の1回目のループにおいて、説明変数p10とp20を除外することが改善をもたらした。2回目のループでは改善がなく、ステップS 2 1 4～S 2 2 6を終了するが、ステップS 2 2 7の判断により再度S 1 1 2に戻る。次に、追加処理の1回目のループにおいて、p13の追加だけが改善をもたらしたが、続く除外処理の1回目のループでは、改善がなかった。もう一度ステップS 1 1 2に戻り、ステップS 1 1 4～S 1 2 6およびステップS 2 1 4～S 2 2 6では改善がなくなったのを確認して、処理を終了した。こうして選択された説明変数は、p3、p6、p12、p13、p16およびp18の5個であり、PRESS=0.048となった。詳細は表4のとおりである。

表4は、表2の段階10まで処理が進んだ時点での部分最小自乗法のフィットならびにリープ・ワン・アウト予測状況を示す。

表4 表2の段階10での処理結果

#	実測値	モデル予測		リープ・ワン・アウト予測	
		計算値	誤差	計算値	誤差
1	0.713	0.771	-0.058	0.663	0.050
2	0.133	-0.013	0.146	0.041	0.092
3	0.545	0.610	-0.065	0.595	-0.050
4	0.752	0.524	0.228	0.380	0.372
5	0.900	0.696	0.205	0.543	0.357
6	0.455	0.591	-0.137	0.623	-0.168
7	0.427	0.638	-0.211	0.696	-0.269
8	0.042	0.189	-0.147	0.268	-0.226
9	0.935	0.841	0.094	0.756	0.179
10	0.154	0.209	-0.055	0.294	-0.140
		0.093	0.022	0.048	
		$R^2=0.765$		$Q^2=0.482$	

なお、説明変数の数が多い時に強いとされる部分最小自乗法であるが、p20～p2の全てを説明変数として採用した場合には、表5に示すように、PRESS=0.124となった。すなわち、リーブ・ワン・アウト処理は、平均値からの誤差(0.093)よりも悪い成績をもたらす。

表5 全ての説明変数を採用した場合の処理結果

#	モデル予測			リーブワンアウト予測	
	実測値	計算値	誤差	計算値	誤差
1	0.713	0.712	0.001	0.527	0.186
2	0.133	-0.073	0.206	0.222	-0.090
3	0.545	0.561	-0.016	0.538	0.007
4	0.752	0.656	0.096	0.351	0.402
5	0.900	0.691	0.209	0.432	0.469
6	0.455	0.519	-0.064	0.562	-0.107
7	0.427	0.583	-0.156	0.629	-0.203
8	0.042	0.430	-0.388	0.724	-0.682
9	0.935	0.794	0.140	0.480	0.454
10	0.154	0.182	-0.029	0.457	-0.303

0.093

0.029

0.124

 $R^2=0.684$ $Q^2=-0.330$

実施例.

次に、実施例を挙げて本発明をさらに詳細に説明するが、本発明はこれらの例によって何ら限定されるものではない。

実施例1： 部分最小自乗法の交差検証成績を考慮した特徴抽出によるDLB

CL患者のデータ解析.

P. O. Brownらのホームページ (<http://llmpp.nih.gov/lymphoma/>) より入手した28名のDLBCL (リンパ腫) 患者のデータを、20名のデータからなる訓練集合と8名のデータからなるテスト集合に分けた。目的変数に生存月数を採用し、説明変数には18432スポットのうち、28データにおいてch1、ch2ともに正の数となる12832スポットの $\log(ch1/ch2)$ 値を採用した。

訓練集合において部分最小自乗法 (PLS) のモデル決定を試みた。12832変数全てを用いて部分最小自乗法の解析をしたところ、リーブ・ワン・アウト予測は有意($Q^2 > 0.5$)にはならなかった。次にリーブ・ワン・アウト予測誤差が最小になるように説明変数を段階的に1つずつ増減した。モデル構成手法としては前述の第3のモデル構成手法において説明変数の追加及び除外の順番並びにリーブ・ワン・アウト処理におけるサンプルの除外の順番が異なるほかは同様な方法を用いた。すなわち、どの説明変数も選択されていない状態を初期状態とする(S112)。次に、最後(m番目)の説明変数から最初(1番目)の説明変数までの未だ選択されていない説明変数ごとに逐次、その説明変数を選択した場合に交差検証成績が改善するかどうかを、リーブ・ワン・アウト処理(ここでは、最後(n番目)のサンプルから最初(1番目)のサンプルを逐次除外した)を用いた交差検証成績評価ステップを繰り返しながら判定し、改善する場合にはその説明変数を追加選択し、そのような改善と追加がなくなるまでm番目の説明変数から上記逐次判定操作を繰り返す(S114~S126)。次に、最後(m番目)の説明変数から最初(1番目)の説明変数までの選択されている説明変数ごとに逐次、その説明変数を除外した場合に交差検証成績が改善するかどうかを、リーブ・ワン・アウト処理(ここでも最後(n番目)のサンプルから逐次除外した)を用いた交差検証成績評価ステップを繰り返しながら判定し、改善する場合にはその説明変数を除外し、そのような改善と除外がなくなるまで、最後(m番目)の説明変数から上記逐次判定操作を繰り返す(S214~S226)。その結果、有意なモデル($R^2 = 0.988$ 、 $Q^2 = 0.895$ 、 $NP = 342$)を得た。図9は、このデータについての最小自乗法成績を示す。図9において、ひし形(fit)は訓練集合のデータ(20人)を示し、三角(cv)は、それらについての交差検証成績

績のデータを示す。また、四角 (test) はテスト集合のデータ (8 人) を示す。
得られた部分最小自乗法モデルは、テスト集合のうち、4/8 をきわめて良好に、
また 1/8 を良好に予測するものであった。

5 なお、上述の多変量解析によるデータ解析では、扱ったサンプルは DNA チップを用いて得たデータであった。しかし、このデータ解析は、DNA チップを用いて得たデータに限定されるものではなく、蛋白質発現量、細胞内物質の量などのデータに対しても有用であろうことは容易に推測されることである。

10 以下の実施例 2 ~ 7 では、部分最小自乗法を用いて選抜した少ない個数の説明変数について、通常の統計的手法または多変量解析手法 (比例ハザード法、重回帰分析、適応最小自乗法、ロジスティック回帰分析法、線型判別分析法など) を適用する。

実施例 2 : 部分最小自乗法の交差検証成績を考慮した特徴抽出と比例ハザード解析による 240 名の DLBCL 患者の生存時間解析。

15 Rosenwald らが Web 上 (<http://l1mmp.nih.gov/DLBCL/>) で公開している 240 名の DLBCL (びまん性大細胞型 B リンパ腫) のデータセットをダウンロードして用いた。全データを訓練集合として利用した。スポットパターンで x_1 または x_2 が 0 となるものを除いた 7399 スポットについて $\log(x_1/x_2)$ を計算して説明変数とした。本実施例では実施例 1 と異なり、生存時間として観測打ち切り時間と死亡時間とが混在していることを考慮してカプラン・マイヤー (Kaplan-Meier) 法による生命表を適用して事象発生時点での生存確率 (P_{KM}) を求め、ロジット変換 ($\log(P_{KM}/1-P_{KM})$) した値を目的変数とした。カプラン・マイヤー法による生命表は集団としての生存確率を示すが、ここでは、個人 j を含む集団としての事象発生時点での残存確率 (変化の発生しなかったものが残存する確率) を個人 j の事象発生時点での残存時間に読み代えるという新規な考え方をを用いている。
25 また、この確率をロジット変換して、変化の発生傾向を表現するロジット値に変換して、目的変数とした。訓練集合内の交差検証はリーブ・ワン・アウト法によって行ない、 $PRESS \times 1.02^N$ が小さくなるようにパラメータを逐次取捨選択して部分最小自乗法モデルを得た。ここで、交差検証成績 ($CV = -PRESS$) の代わり

に、少なくとも交差検証成績を独立変数として持つ関数の1つである関数-PRESS $\times 1.02^{N P}$ を改善して部分最小自乗法モデルを得た。ここでPRESSはリーブ・ワン・アウト予測の残差自乗和であり、NPは、選択された説明変数の数である。

図7のフロー中の交差検証成績CVを $-\text{PRESS} \times 1.02^{N P}$ と読み換えて、処理を
5 実行することにより、下記の19個の遺伝子の発現が説明変数として選抜された。
ここでdata IDはWebデータ元でのID番号を示す。またACCESSIONはGenBankのアクセシ
ョン番号であり、アクセション番号の無い行はデータ元でのみ明らかとなっ
ている遺伝子 (Unknown) ないしESTであり、論文記載の方法によって入手するこ
とができる。

10
ACCESSION data ID comment
U03398 #(27876) tumor necrosis factor (ligand)
superfamily, member 9
M65066 #(27394) protein kinase, cAMP-dependent,
15 regulatory, type I, beta
-- #(27104) (Unknown)
AK001546#(25048) Homo sapiens cDNA FLJ10684 fis, clone
NT2RP3000220
-- #(31372) (Unknown)
20 U15085 #(28178) major histocompatibility complex,
class II, DM beta
BC003536#(24983) hypothetical protein MGC10796
-- #(16113) (Unknown)
M23452 #(16822) small inducible cytokine A3
25 #(24433) (Unknown)
X00437 #(27480) T cell receptor beta locus
U12979 #(24377) activated RNA polymerase II
transcription cofactor 4
X52479 #(17773) protein kinase C, alpha

H96306 #(16578) bone marrow stromal cell antigen 1
 U70426 #(19255) regulator of G-protein signaling 16
 AA830781#(33358) EST
 AA804793#(25022) EST
 5 H57330 #(26383) EST
 S69790 #(27184) WAS protein family, member 3

これらの遺伝子の発現を説明変数の候補として比例ハザード(hazard)解析を試みた。比例ハザード法とは、生存率の解析に時間を考慮した統計的手法である。
 10 解析の実行はプログラムパッケージ JMP (JMP Sales SAS Campus Drive Cary, NC 27513 USA)を用いて行なった。変数削除基準として $P \geq 0.05$ を採用した変数減少法によって更に絞り込んだ結果、14 遺伝子の発現からなる以下の比例ハザード式が得られた。ここでGenbank (ジーンバンク) のアクセション番号ないし data IDで示される各項は、各遺伝子の $\log(x_1/x_2)$ 値であり、またPは統計
 15 的な有意性が成り立たない危険率である。この式の右辺から求められるハザード値(hazard)が大きいほど、死亡傾向が大きい。

$$\begin{aligned} \text{hazard} = & 0.370 \text{ \#(27104) } + 0.589 \text{ AK001546 } - 0.366 \text{ \#(31372) } - 0.276 \text{ U15085 } \\ & - 0.307 \text{ \#(16113) } + 0.409 \text{ M23452 } - 0.350 \text{ \#(24433) } - 0.297 \text{ X00437 } \\ & + 0.321 \text{ U12979 } - 0.585 \text{ X52479 } - 0.457 \text{ U70426 } + 0.561 \text{ AA830781 } \\ & - 0.430 \text{ H57330 } + 0.433 \text{ S69790 } \end{aligned}$$

$$P < 0.0001$$

Rosenwaldらは、単相関の比例ハザード解析を行なって、5群(17 遺伝子)の診断指標を選抜している。図10に、本実施例で得られたハザード値(Hazard、
 図中 Hazard (pls(14))と示した)とRosenwaldらの診断指標がどの程度、生存時
 25 間を説明できているかを比較した。Rosenwaldらの5群のパラメータを同時に用いた比例ハザード式ではProlifirationパラメータが $P > 0.05$ で統計的に有意でないなどの問題を有していたため、これを除く4群のパラメータを同時に含めたハザード値も比較のために掲載した(図中 Hazard (Rosenwald/4para)と示した)。ここで、菱形は死亡した人または打ち切った人のデータを示し、四角は生存してい

る人のデータを示す。

これらの診断指標のうち、本実施例で求めたハザード値と生存時間との相関は際立って明白である。即ちハザード値は生存時間につれて減衰しており、大きなハザード値の患者は長く生きることが出来ないことが示されている。一方、

5 Rosenwaldらの指標はいずれも生存時間を診断するには不十分なものである。数百、数千という数のパラメータの中から効率的に最適のパラメータセットを見出すことは比例ハザード解析だけではできないことである。しかし以上のように Kaplan・マイヤー法、ロジット変換、部分最小自乗法の交差検証成績を考慮した特徴抽出、比例ハザード解析を組み合わせることで、従来に無い、有効な診断指標を得ることができた。統計学的に異質なモデルをこのように組み合わせること

10 によってこのような良好な結果が得られたことは意外でもあり、興味深いことであった。患者の生存時間を予測することは、QOLを含めた治療計画や人生設計などを判断する上で重要な情報をもたらすものであり、本実施例で求められた診断モデルは社会的に価値のあるものである。

15 また、変数削除基準として $P \geq 0.001$ を採用した変数減少法によって更に絞り込むと、6 遺伝子の発現からなる以下の比例ハザード式が得られた。このように、変数削除基準を変えることにより、選択される説明変数の数を制御できる。

$$\begin{aligned} \text{hazard} = & -0.426 \text{ U15085} + 0.350 \text{ M23452} - 0.521 \text{ X52479} \\ & - 0.450 \text{ U70426} - 0.586 \text{ H57330} + 0.476 \text{ S69790} \end{aligned}$$

20 図 1 1 は、右辺を計算して求められるハザード値を縦軸とし、生存時間を横軸としたプロットを示す。図 1 0 と同様に、図 1 1 において、菱形は死亡した人または打ち切った人のデータを示し、四角は生存している人のデータを示す。

実施例 3 : 部分最小自乗法の交差検証成績を考慮した特徴抽出と比例ハザード解析による 40 名の乳癌患者の生存時間解析。

25 SorleらがWeb上 (http://genome-www.stanford.edu/breast_cancer/mopo_clinical/) で公開している乳癌患者のデータセットをダウンロードして用いた。全データを訓練集合として利用した。データセットの大部分は、タイプ A, B という 2 種類の DNA チップで測定された

それぞれ40名、24名の患者よりなるが、ここではタイプAのデータを用いた。生存時間データより実施例2と同様にロジット値を求め、目的変数とした。説明変数としては、データに欠測のある遺伝子を除いた6891件のLOG_RAT2N_MEAN値を採用した。そして、少なくとも交差検証成績を独立変数として持つ関数の1つである、交差検証成績と説明変数NPの関数 $\text{PRESS} \times 1.13^N$ が小さくなるようにパラメータを逐次取捨選択して部分最小自乗法モデルを得た。図7のフロー中の交差検証成績CVを $-\text{PRESS} \times 1.13^N$ と読み換えて、処理を実行することにより、下記の10個の遺伝子の発現が説明変数として選抜された。

ACCESSION comment

- 10 AA406242 (guanosine monophosphate reductase)
 AA598572 (spleen tyrosine kinase)
 H73335 (Homo sapiens mRNA full length insert cDNA clone EUROIMAGE
 980547)
 W84753 (Homo sapiens cDNA FLJ13510 fis, clone PLACE1005146)
 15 AA703058 (myeloperoxidase)
 N71160 (cytochrome c oxidase subunit Vib)
 AA453345 (a protein tyrosine kinase)
 AA054669 (Homo sapiens, clone IMAGE:3611719, mRNA, partial cds)
 N32820 (ESTs, Weakly similar to ALU1_HUMAN ALU SUBFAMILY J SEQUENCE
 20 CONTAMINATION WARNING ENTRY [H. sapiens])
 R05667 (suppressor of potassium transport defect 3)

これらを説明変数の候補として、比例ハザード解析において変数削除基準として $P \geq 0.05$ を採用した変数減少法を試み、7遺伝子の発現からなる以下の比例ハザード式が得られた。ここでアクセッション番号で示される各項はそれぞれの遺伝子のLOG_RAT2N_MEANである。

$$\text{hazard} = -0.821 \text{ AA406242} + 1.556 \text{ AA598572} - 1.074 \text{ H73335} + 1.418 \text{ W84753} \\ - 1.290 \text{ AA703058} + 2.182 \text{ N71160} + 0.828 \text{ AA453345}$$

$P < 0.0001$ 変数の $P < 0.05$

図12に、右边を計算して求められるハザード値を縦軸とし、生存時間を横軸

としたプロットを示す。ここでもハザード値が優れた診断指標となることが示されている。図 1 2において、菱形は死亡した人または打ち切った人のデータを示し、四角は生存している人のデータを示す。

変数削除基準として $P \geq 0.001$ を採用した変数減少法によって更に絞り込んだ。

5 これにより、3 遺伝子の発現からなる以下の比例ハザード式が得られた。このように、変数削除基準を変えることにより、説明変数の数を制御できた。

$$\text{hazard} = 1.453 \text{ AA598572} - 1.473 \text{ AA703058} + 1.071 \text{ AA453345}$$

図 1 3は、右辺を計算して求められるハザード値を縦軸とし、生存時間を横軸としたプロットを示す。ここで、菱形は死亡した人のデータを示し、四角は生存している人のデータを示す。

実施例 4 : 部分最小自乗法の交差検証成績を考慮した特徴抽出と重回帰分析による 40 名の乳癌患者の再発予測解析。

15 SorleらのDNAチップAで6891 遺伝子の発現が測定された 40 名の患者をデータセットとして用いた。再発の有無を目的変数として、 $\text{PRESS} \times 1.10^N$ が小さくなるようにパラメータを逐次取捨選択して 11 遺伝子の発現からなる部分最小自乗法モデルを得た。

ACCESSION	comment
AA434397	integrin, beta 5
20 T83209	ESTs
N53427	KIAA1628 protein
N29639	cytidine monophosphate-N-acetylneuraminic acid hydroxylase
AA485739	major histocompatibility complex, class II, 25 DR beta 5
AA425861	enoyl Coenzyme A hydratase 1, peroxisomal
H84871	Ste-20 related kinase
T64312	prostate cancer overexpressed gene 1
T59518	solute carrier family 2, (facilitated glucose

transporter) member 8

AA406231 KIAA0381 protein

AA037488 prolactin

- 次に、選抜された遺伝子発現を説明変数とし、再発の有無を目的変数として、
5 通常の多変数解析法の一つである重回帰分析によって判別分析を実行した。解析
の実行はプログラムパッケージ JMP を用いて行なった。変数削除基準として $P \geq 0.15$ を採用した変数減少法によってさらに絞り込んだ結果、10 遺伝子の発現
からなる以下の重回帰式が得られた。この式で計算される OLS 値が正の時は再発
の可能性が高く、負の時は低い。

10
$$\begin{aligned} \text{OLS} = & -0.215 \text{ AA434397} + 0.227 \text{ T83209} - 0.209 \text{ N53427} + 0.139 \text{ N29639} \\ & + 0.165 \text{ AA485739} + 0.133 \text{ AA425861} - 0.084 \text{ H84871} - 0.193 \text{ T64312} \\ & + 0.237 \text{ T59518} + 0.176 \text{ AA037488} - 0.278 \end{aligned}$$

$R^2 = 0.84797$ 、 判別正解率 97.5%

- 上式に含まれる各パラメータをそれぞれ 1 つ用いて判別分析式を作成した場合
15 の P 値及び決定係数を以下の表 6 に示す。

表 6

Accession No.	P value	決定係数(R^2)
AA434397	0.0334	0.090273
T83209	0.0601	0.066005
N53427	0.0004	0.268678
N29639	0.0552	0.069483
AA485739	0.0421	0.080733
AA425861	0.0861	0.05122
H84871	0.087566	0.087566
T64312	0.0004	0.263207
T59518	0.0066	0.157196
AA037488	0.0031	0.187627

単独では有意とはならない ($P > 0.05$) パラメータが 3 つ存在し、また、どのパラ

メータも決定係数が小さい。従って、パラメータを1つずつ吟味するだけでは、上式のような良好な判別式は得られなかった。また数百、数千という数のパラメータの中から効率的に最適のパラメータセットを見出すことは重回帰分析だけではできないことである。しかし、以上のように、部分最小自乗法の交差検証成績を考慮して特徴抽出することにより、従来に無い、有効な診断指標を得ることができた。乳癌の再発可能性を予測することは、QOLを考慮した治療計画を立案し判断するうえで、社会的に求められているところのものである。

実施例5： 部分最小自乗法の交差検証成績を考慮した特徴抽出と適応最小自乗法による40+24名の乳癌患者の再発予測解析。

DNAチップのタイプA(40名)とタイプB(24名)に共通する3448遺伝子に限って解析を試みた。PRESS $\times 1.17^N$ が小さくなるようにパラメータを逐次取捨選択して部分最小自乗法モデルを得た。選抜された遺伝子発現を説明変数とし、適応最小自乗法によって判別分析を実行した結果、次式が得られた。次式で計算されるALS値が0.5より大きいと再発の危険性が存在する。

$$ALS = 0.31 H11482 - 0.29 T64312 - 0.32 AA045340 + 0.01$$

$$R^2 = 0.65, \text{ eps} = 0.13, \text{ 判別正解率 } 90.0\%$$

下記の表7にみるように、H11482は単相関では有意ではなく、他の変数と同時に用いることで初めて把握できたパラメータである。また、表8は、上式を用いてタイプBの患者を予測した結果である。本判別式の感度=81.8%、特異度=53.8%となり、 $\chi^2=3.233$ (5% $P<10\%$)、予測判別正解率=66.7%、という統計的に有意な結果を得た。タイプA、BはDNAチップの構成の相違に基づく測定誤差が存在すると思われるデータであるにもかかわらず、タイプAで訓練したモデルでタイプBの予測に危険率10%以下で成功したことは勇気付けられる結果である。

また、PRESS $\times 1.12^N$ が小さくなるように選んだ場合には、以下の遺伝子の発現を説明変数とする部分最小自乗法モデルを得た。

$$H11482, T64312, R99749, T65211, AA427625, AA455506$$

これらを説明変数の候補として、リーブ・ワン・アウトを指標にして、さらに

絞り込んだ結果、次の判別式を得た。

$$ALS = 0.53 H11482 - 0.31 T64312 - 0.33 R99749 - 0.26 AA455506 + 0.10$$

$$R^2 = 1.00, \text{ eps} = 0.10, \text{ 判別正解率 } 100.0\%$$

パラメータを1つずつ吟味するだけでは、上式のような良好な判別式は得られなかった。また数百、数千という数のパラメータの中から効率的に最適のパラメータセットを見出すことは、適応最小自乗法、ロジスティック回帰分析、その他の判別分析手法だけではできないことである。しかし、以上のように、部分最小自乗法の交差検証成績を考慮して特徴抽出することにより、従来に無い、有効な診断指標を得ることができた。

表7 パラメータの交絡作用

パラメータ	R	Nmis (/40)
H22482	0.361	14
T64312	0.607	8
AA045340	0.572	9
T64312 & AA045340	0.716	6
H11482 & T64312 & AA045340	0.804	4

表8 タイプBの24患者の予測

観察値	予測値	頻度
—	—	7
+	—	2
—	+	6
+	+	9

実施例6： 部分最小自乗法の交差検証成績を考慮した特徴抽出とロジスティック回帰分析法または線型判別分析法による40+24名の乳癌患者の再発予測解析。

実施例5での1つめの適応最小自乗法による解析をロジスティック回帰分析法

に置き換えた場合、次の判別式が得られた。

$$LORA = 7.92 \text{ H11482 } -5.69 \text{ T64312 } -6.41 \text{ AA045340 } -9.73$$

$$R^2 = 0.63, x^2 = 35.00 \text{ (P<0.0001), 判別正解率 } 90.0\%$$

右辺で求められるLORA値が正の場合には再発の危険性が存在する。係数の
5 比率や相関係数は実施例5の適応最小自乗法の場合と異なるものの、各患者の識別結果は全く同一であった。またタイプBの患者を予測した結果も表7と同じになった。

次に、実施例5での適応最小自乗法による解析を線型判別分析に置き換えて解析して、次の判別式が得られた。

$$10 \quad LDA = 2.45 \text{ H11482 } -2.35 \text{ T64312 } -2.56 \text{ AA045340 } -4.03$$

$$\text{判別正解率 } 80.0\%$$

右辺で求められるLDA値が正の場合には再発の危険性が存在する。係数の比率や相関係数は、実施例5の適応最小自乗法の場合と異なり、各患者の識別結果も若干異なったが、概ね同一であった。また、タイプBの患者を予測した結果も
15 表7と同じになった。

以上の実施例4, 5, 6では、乳癌の再発の有無を目的変数としている。したがって、部分最小自乗法の交差検証成績を考慮して特徴抽出する方法が、目的変数が名義尺度や順序尺度などのデータである場合にも有効であることが示された。なお、名義尺度とは、対象（サンプル）をある分類に属するかどうかを測り分け
20 るときの分類で、分類の間に大小や順序はない。また、順序尺度とは、対象の特定の分類について測り分けるときの分類であり、分類の間に大小、高低といった順序がある。

実施例7： 部分最小自乗法の交差検証成績を考慮した特徴抽出と比例ハザード解析による40名の乳癌患者の再発時間解析。
25

実施例4と同じデータを用いて、再発の時系列データを基に実施例2と同様の方法で求めたロジット値を目的変数として、 $\text{PRESS} \times 1.15^P$ が小さくなるようにパラメータを逐次取捨選択して9遺伝子の発現からなる部分最小自乗法モデルを得た。これらの遺伝子発現の測定値を説明変数として比例ハザード解析において

変数削除基準として $P \geq 0.05$ を採用した変数減少法を試み、8遺伝子からなる、以下の比例ハザード式が得られた。

$$\begin{aligned} \text{hazard} = & 1.122 \text{ AA448641} -1.781 \text{ R78516} -1.434 \text{ R05934} +2.165 \text{ W84753} \\ & -1.923 \text{ AA629838} +2.665 \text{ H08581} +1.875 \text{ AA045730} +1.269 \text{ AI250654} \end{aligned}$$

$$P < 0.0001$$

図14は、右辺を計算して求められるハザード値を縦軸とし、再発時間を横軸としたプロットを示す。ここで、菱形は再発しない人のデータを示し、四角は再発している人のデータを示す。ここでもハザード値が優れた診断指標となっており、生存時間に限らず、時間とともに確率的に発生する生体の状態の変化を解析する手法として、本発明の手法が有効であることが示されている。

変数削除基準として $P \geq 0.005$ を採用した変数減少法によって更に絞り込んだ場合には、4遺伝子の発現からなる以下の比例ハザード式が得られた。

$$\text{hazard} = 1.559 \text{ W84753} +2.265 \text{ H08581} +1.473 \text{ AA045730} +1.237 \text{ AI250654}$$

図15は、右辺を計算して求められるハザード値を縦軸とし、再発時間を横軸としたプロットを示す。ここで、菱形は再発しない人のデータを示し、四角は再発している人のデータを示す。

実施例8： Genbankアクセッション番号H11482、T64312、AA045340を含む乳癌再発性診断用DNAチップの作成と測定。

実験医学別冊「ゲノム機能研究プロトコール」(ISBN4-89706-932-7 C3047) p34-38記載の関直彦、永杉友美、東孝典、吉川勉、鈴木収、村松正明らの方法に準じてDNAチップの作成と測定を行なった。Genbankアクセッション番号H11482、T64312、AA045340のcDNAを用いた。

プローブ用の各PCR産物をエタノール(和光純薬, Cat#057-00456)で沈殿させ、 $2 \mu\text{g}/\mu\text{l}$ となるようにDDWで調整する。ニトロセルロース(GibcoBRL Cat#41051-012) 4 mg/ml のDMSO溶液を等量加え、よく混和させて 100°C で5分間熱変性を行ない、氷上で急冷する。次いで室温に戻し、DNAスポットターSPBIO2000(日立ソフトエンジニアリング)を用いてカルボジイミドスライドガラス(日清紡)へのスポッティングを速やかに行なう。スポットの乾燥を確認し、

Ultraviolet crosslinker(アマシヤムファルマシアバイオテック社)を用いて60mJ/cm²で紫外クロスリンク処理を行ない、ガラスラックに立てて室温保存する。

3%BSA、0.2M NaCl、0.1M Tris(PH 7.5)、0.05% Triton X-100よりなるブロッキング液に上記マイクロアレイを浸け、約30分間放置する。次いで、ガラスに付着している溶液をよく切り、37℃で乾燥させる。TEバッファー(PH 8.0, ニッポンジーン Cat #316-90025)で3回軽く洗い、プレートホルダーに入れて軽く遠心(1000 rpm, 1分間)して余分な水分を除去する。

次に、乳腺正常株SV-40及び乳癌細胞株MCF-7、MDA-MB-468又はT-47-Dの各細胞液より、TRIZOL (GibcoBRL, Cat#15596-018)、Oligotex dT30<Super> (TaKaRa, Cat#W9021A)を用いてマニュアルに従って、mRNAを精製する。2μgのmRNAをDEPC処理した6.4μlのDDWに溶かし、Oligo dTプライマー 9μl、5 × SuperScript IIバッファー(GibcoBRL, Cat#18089-011) 6μl、DTT (SuperScriptの付属) 3μl、50 × dNTP 0.6μl、Cy3-dUTP(アマシヤムファルマシアバイオテック Cat# PA53022)又はCy5-dUTP (アマシヤムファルマシアバイオテック Cat# PA55022) 3μl、SuperScript II 2μlよりなる溶液を加え、42℃で2時間反応させる。途中1時間経過時点で、SuperScript IIを1μlを追加する。1. 5μlアルカリバッファー(1N NaOH / 20nM EDTA)を加え、65℃で10分間反応させ、TEバッファーを270μl、1N HClを1.5μl加えて、Cy3, Cy5ラベルの反応液を2つまとめて1本のMicrocon-YM-30 (Millipore/Amicon, Cat#42410)に移す。10,000rpmで上のカップに残る液量が約10μlになるまで遠心を続け、カップを通りぬける液を別のチューブに移し替え、その後、上のカップにTE バッファー500μl、Human Cot-1 DNA (GibcoBRL Cat#15279-011) 20μgを加え、再び液量が10μl以下になるまで遠心を続ける。3,000rpmで3分間遠心し、蛍光標識したDNAを回収する。DDWとyeast RNA (Sigma, Cat#R7125) 50μg、poly(A) (ロッシユダイアグノスティクス, Cat#108 626) 50μgを加えて20μlにし、PCR用のチューブに移し換え、さらに4.25μl 20 × SSC (GibcoBRL, Cat#15553-035)と0.75μl 10% SDS (GibcoBRL, Cat#15553-035)を加え、PCR用の機器で100℃、1分間熱変性させ、次いで、室温で30分間放置して、ゆっくり冷却する。

蛍光標識したDNAの全量をカバーガラスにのせ、泡が入らないように注意しながら前記マイクロアレイにかぶせ、水で濡らしたキムタオルを底に敷いたハイブリダイゼーションチェンバーに入れて密閉する。毎分2～4サイクルで軽く振とうさせながら、65℃で一晩ハイブリダイズする。ハイブリダイゼーションチェンバーからマイクロアレイを取り出し、カバーガラスが載ったままの状態です
5 静かに2×SSC/0.1% SDS溶液中に入れ、5分間シェイキングし、カバーガラスが自然にはがれるのを待つ。カバーガラスがはがれたところでマイクロアレイをスライドガラスラックに入れ、もう一度2×SSC/0.1% SDS溶液中で5分間軽く振とうして洗う。さらに0.2×SSC/0.1% SDS
10 40℃で5分間2回洗い、0.2×SSCでリンスする。マイクロアレイを別の乾いたプレパラートケースに移し、マイクロタイタープレート用の遠心機で軽く遠心して(1000 rpm, 1分室温)マイクロアレイ上の水分を除く。そして、ScanArray4000 (GSI luminonics社)でシグナルを読み込み、解析ソフトにはQuant Array (GSI luminonics社)およびChip Space(日立ソフトウェアエンジニアリング)を用いる。
15

実施例9： 遺伝的アルゴリズムによる部分最小自乗法モデルの最適化。

実施例4で用いたSorleらのDNAチップAで6891遺伝子の発現が測定された40名の患者をデータセットとして用いた。遺伝的アルゴリズムは、たとえば、伊庭斉志；「遺伝的アルゴリズムの基礎」(オーム社(1994))に説明されている。前記データを用い、遺伝的アルゴリズムによる説明変数選択を行なった。以下において「」で区切られた用語は遺伝的アルゴリズムで通常用いられる専門用語であり、特に必要な場合には解説を加えている。「適合度」(fitness)にはPRESS×1.01⁹⁰を採用した。各「個体」の「遺伝型」は説明変数を採用する場合には1、採用しない場合には0をとる数列{b1, b2, b3, ...}とした。
20
25

個体集合のサイズを100個とし、初期の個体の「遺伝型」(GTYPE)は、平均でmin_of(Ns, Ng, 300)/2個の説明変数が採用となるように乱数を用いて準備した。ここでNsはサンプル数(患者数)、Ngは説明変数の候補の数、300は実装の都合上設定された定数である。

集合よりランダムに2つの個体を選抜し、「遺伝型」の「一様交叉」を行なったものの一方を新しい「個体」とした。即ち、「各遺伝子座」ごとに1/2の確率でいずれかの「親個体」の数値(0または1)を選びそれを代入したものを新しい「個体」とした。続いて新しい「個体」の「各遺伝子座」毎に、1の場合(説明変数が採用されている場合)には1.1/採用された説明変数の数の確率で、0の場合(採用されていない場合)には1.1/採用されていない説明変数候補の数の確率で、 $0 \leftrightarrow 1$ を反転させた。

上述の「交叉・突然変異オペレーション」によって準備された新しい「個体」の「適合度」と、ランダムに選抜された「トーナメント相手」となる集合中の「個体」の「適合度」とを比較し、新しい「個体」の適合度が勝った場合には0.75の確率で、劣った場合には0.25の確率で「個体」の置き換えを行なった。ただし、「トーナメント相手」が集合中の最適解のものである場合には置き換えを禁止するという「エリート戦略」を採用した。

以上の「交叉」→「突然変異」→「選抜」サイクルを繰り返して最適化を行なった。ここではサイクル数を集合サイズで割ったものを「世代数」とする。最大「世代数」の初期値を100とし、新しい最適解が見出されるたびに最大「世代数」を10増加させながら、実行「世代数」が最大「世代数」に至るまでサイクルを繰り返した。

以上の初期集合の準備～最適化の繰り返しおよび終了にいたる一連の処理を一回のラン(run)とし、15回のランを行なった。図16は、15回のランにおける最適化の様子をまとめている。最良の結果は25個の説明変数を用いたものである。

実施例10： 階層型人工ニューラルネットワーク(MLP)によるモデル構築。

実施例5の乳癌患者の再発性判別解析において、DNAチップtype A(40名)とtype B(24名)に共通する3448遺伝子より、PRESS $\times 1.17^{100}$ が小さくなるようにしてPLS-CVで特徴抽出された3つの説明変数を用いた。

解析方法について説明すると、MLPは3層とし、中間層(tk)において一度だけシグモイド変換を行なう構造とし、図17の4つのトポロジーを試みた。ネッ

トワークの重みの学習はBack propagation(逆伝播)アルゴリズムによって行なった。中間層(tk)において一度だけシグモイド変換を行なう3層MLPを用いた。

$$s_{ik} = \sum_j w_{kj} \cdot P_{ij}$$

$$t_{ik} = 1 / (1 + \exp\{-s_{ik}\})$$

$$5 \quad y_i = \sum_k v_k \cdot t_{ik}$$

ネットワークトポロジーIおよびトポロジーIIbの結果は以下のとおりであった。なお、トポロジーIIa及びトポロジーIIcは、トポロジーIIbに劣るものであった。トポロジーI:

$$y = 0.76 - 1.77 t_1$$

$$10 \quad s_1 = -12.48 - 42.89 H11482 + 39.38 AA045340 + 29.65 T64312$$

$$R^2 = 0.717 \quad Q^2 = 0.142$$

トポロジーIIb:

$$y = 1.19 - 0.86 t_1 - 1.43 t_2$$

$$t_1 = 2.65 + 18.25 AA045340$$

$$15 \quad t_2 = -0.40 - 2.29 H11482 + 3.55 T64312$$

$$R^2 = 0.626 \quad Q^2 = 0.416$$

実施例11: 潜在変数を用いた比例ハザードモデルの構築。

20 実施例3のPLS-CV法で選抜された10遺伝子の発現量を説明変数とし、目的変数として生存確率のlogit値を用いてPLSの解析過程で作成される潜在変数を1個抽出した。その抽出した潜在変数を説明変数にして比例ハザードモデルによる解析を試みた結果、作成された式は $P \leq 0.0001$ で有意となった。図18に右辺を計算して得られるハザード値を縦軸とし、生存時間を横軸にしたプロットを示す。

25 本技術で得られたハザード式の予測の性能を評価するために、用いた40例の中から1例を除外し、残りの39例のデータを用いてハザード式を作成し、除外した1例のハザード値を予測した。39例からのハザード式によって予測した値と40例からのハザード式からの計算値をプロットした図19より、本技術はハザード値の予測において良好な成績を示した。

発明の効果について以下に説明すると、生体の状態と複数の遺伝子発現の量および／または細胞内物質の量との相関モデルを決定するとき、説明変数の選択と交差検証法とを用いて変数を絞り込むことができる。これにより、良好でかつ予測力のある多変量解析モデル（相関モデル）が得られる。特に遺伝子発現の量のように、説明変数の数がたとえば1000以上と膨大な場合に有用である。変数の数を少なくすることにより、病気や生体现象の背後で働いている重要な遺伝子やメカニズムを推定／特定でき、理解が深まる。また、重要な遺伝子産物や細胞内物質だけに絞った廉価な診断用材料（DNAチップ、DNA含有ベクター、抗体チップなど）を設計し、提供できる。

また、時間とともに確率的に発生する生体の状態の変化から導出された量を目的変数として用いて、時間とともに確率的に発生する生体の状態の変化と複数の遺伝子発現の量および／または細胞内物質の量との相関モデルを決定できる。

また、部分最小自乗法を用いて説明変数の個数を少なくすると、通常の統計的手法または多変量解析手法が適用可能になる。

請 求 の 範 囲

1. 生体の状態または時間とともに確率的に発生する生体の状態の変化を目的変数とし、複数の遺伝子発現の量および／または細胞内物質の量を説明変数とする
5 相関モデルを決定するデータ解析装置であって、

生体の状態或いはそれを導出するデータまたは時間とともに確率的に発生する生体の状態の変化に関するデータと、複数の遺伝子発現の量および／または細胞内物質の量からなるサンプルの集合を入力する入力手段と、

(1)説明変数を選択する選択手段と、

10 (2)部分最小自乗法を実行して交差検証成績を計算する計算手段または前記生体の状態の変化に関するデータにカプラン・マイヤー法又はカトラー・エデラー法による生命表を適用して変化の発生しなかったものの確率を計算して得られた確率を、仮定した分布に基づいた変換または仮定を前提としない変換をし、該変換結果を目的変数とする部分最小自乗法を実行して交差検証成績を計算する計算手
15 段と、

(3)前記(2)の計算手段の結果を評価し、説明変数の採用、不採用を判定する評価判定手段とを有し、

(4)前記(1)の選択手段と前記(2)の計算手段と前記(3)の評価判定手段とを実行して部分最小自乗法モデルの少なくとも交差検証成績を独立変数として持つ関数を改善し続けて部分最小自乗法モデルを決定する決定手段とからなることを
20 特徴とするデータ解析装置。

2. 目的変数が生体の状態であって、前記入力手段で入力するデータが生体の状態或いはそれを導出するデータであって、前記(2)の計算手段が部分最小自乗法を実行して交差検証成績を計算する計算手段であることを特徴とする請求項1
25 に記載のデータ解析装置。

3. 目的変数が時間とともに確率的に発生する生体の状態の変化であって、前記入力手段で入力するデータが時間とともに確率的に発生する生体の状態の変化に関するデータであって、前記(2)の計算手段が前記生体の状態の変化に関するデータにカプラン・マイヤー法又はカトラー・エデラー法による生命表を適用し

て変化の発生しなかったものの確率を計算して得られた確率を、仮定した分布に基づいた変換または仮定を前提としない変換をし、該変換結果を目的変数とする部分最小自乗法を実行して交差検証成績を計算する計算手段であることを特徴とする請求項1に記載のデータ解析装置。

5 4. さらに、前記の決定手段にて決定された部分最小自乗法モデルに採用されている説明変数又は該モデルの潜在変数を用い、統計的手法又は多変量解析手法によるモデルを構築する最終モデル決定手段を備えることを特徴とする請求項1、2又は3に記載のデータ解析装置。

10 5. 前記の選択手段において、説明変数を逐次取捨選択することを特徴とする請求項1～4のいずれかに記載のデータ解析装置。

6. 前記の選択手段において、遺伝的アルゴリズムを用いて説明変数を選択することを特徴とする請求項1～4のいずれかに記載のデータ解析装置。

15 7. 前記の計算手段において、1個のサンプルを逐次除外して部分最小自乗法を実行して交差検証成績を計算することを特徴とする請求項1～6のいずれかに記載のデータ解析装置。

8. 前記の計算手段において、複数のサンプルを逐次除外して部分最小自乗法を実行して交差検証成績を計算することを特徴とする請求項1～6のいずれかに記載のデータ解析装置。

20 9. 前記計算手段において、各計算において除外したサンプルの遺伝子発現から予測される生体の状態を示す目的変数値と、前記除外したサンプルの生体の状態を示す目的変数値との誤差の代表値を求め、交差検証成績の指標として当該誤差を用いることを特徴とする請求項7又は8に記載のデータ解析装置。

10 10. 前記関数が交差検証成績であることを特徴とする請求項1～9のいずれかに記載のデータ解析装置。

25 11. 前記関数が交差検証成績と選択された説明変数の個数との関数であることを特徴とする請求項1～9のいずれかに記載のデータ解析装置。

12. 前記の決定手段において、少なくとも交差検証成績を独立変数として持つ関数を改善しながら評価判定を繰り返すことを特徴とする請求項5に記載のデータ解析装置。

1 3. 前記(1)の選択手段と前記(2)の計算手段とを複数のコンピュータで実行させることを特徴とする請求項1～12のいずれかに記載のデータ解析装置。

1 4. 請求項1、2、3又は4で決定された相関モデル及び予測対象のサンプルについて当該モデルにおいて採用された説明変数を入力する入力手段と、入力された該説明変数に基づいて該サンプルの生体の状態を予測判定する予測判定手段からなることを特徴とするデータ解析装置。

1 5. 生体の状態を名義尺度、順序尺度或いは連続量で表現する目的変数とする請求項2に記載のデータ解析装置。

1 6. 最終モデル決定手段が用いる前記の統計的手法又は多変量解析手法が、比例ハザード法又はパラメトリックな分布にあてはめた回帰分析法であることを特徴とする請求項2又は4に記載のデータ解析装置。

1 7. 生体の状態または時間とともに確率的に発生する生体の状態の変化を目的変数とし、複数の遺伝子発現の量および／または細胞内物質の量を説明変数とする相関モデルを決定するデータ解析方法であって、

生体の状態或いはそれを導出するデータまたは時間とともに確率的に発生する生体の状態の変化に関するデータと、複数の遺伝子発現の量および／または細胞内物質の量からなるサンプルの集合を入力する入力ステップと、

(1)説明変数を選択する選択ステップと、

(2)部分最小自乗法を実行して交差検証成績を計算する計算ステップまたは前記生体の状態の変化に関するデータにカプラン・マイヤー法又はカトラー・エデラ一法による生命表を適用して変化の発生しなかったものの確率を計算して得られた確率を、仮定した分布に基づいた変換または仮定を前提としない変換をし、該変換結果を目的変数とする部分最小自乗法を実行して交差検証成績を計算する計算ステップと、

(3)前記(2)の計算ステップの結果を評価し、説明変数の採用、不採用を判定する評価判定ステップとを有し、

(4)前記(1)の選択ステップと前記(2)の計算ステップと前記(3)の評価判定ステップとを実行して部分最小自乗法モデルの少なくとも交差検証成績を独立変数として持つ関数を改善し続けて部分最小自乗法モデルを決定する決定ステップ

とからなることを特徴とするデータ解析方法。

18. 目的変数が生体の状態であって、前記入力ステップで入力するデータが生体の状態或いはそれを導出するデータであって、前記(2)の計算ステップが部分最小自乗法を実行して交差検証成績を計算する計算ステップであることを特徴とする請求項17に記載のデータ解析方法。

19. 目的変数が時間とともに確率的に発生する生体の状態の変化であって、前記入力ステップで入力するデータが時間とともに確率的に発生する生体の状態の変化に関するデータであって、前記(2)の計算ステップが前記生体の状態の変化に関するデータにカプラン・マイヤー法又はカトラー・エデラー法による生命表を適用して変化の発生しなかったものの確率を計算して得られた確率を、仮定した分布に基づいた変換または仮定を前提としない変換をし、該変換結果を目的変数とする部分最小自乗法を実行して交差検証成績を計算する計算ステップであることを特徴とする請求項17に記載のデータ解析方法。

20. さらに、前記の決定ステップにて決定された部分最小自乗法モデルに採用されている説明変数又は該モデルの潜在変数を用い、統計的手法又は多変量解析手法によるモデルを構築する最終モデル決定ステップを備えることを特徴とする請求項17、18又は19に記載のデータ解析方法。

21. 前記の選択ステップにおいて、説明変数を逐次取捨選択することを特徴とする請求項17～20のいずれかに記載のデータ解析方法。

22. 前記の選択ステップにおいて、遺伝的アルゴリズムを用いて説明変数を選択することを特徴とする請求項17～20のいずれかに記載のデータ解析方法。

23. 前記の計算ステップにおいて、1個のサンプルを逐次除外して部分最小自乗法を実行して交差検証成績を計算することを特徴とする請求項17～22のいずれかに記載のデータ解析方法。

24. 前記の計算ステップにおいて、複数のサンプルを逐次除外して部分最小自乗法を実行して交差検証成績を計算することを特徴とする請求項17～22のいずれかに記載のデータ解析方法。

25. 前記計算ステップにおいて、各計算において除外したサンプルの遺伝子発現から予測される生体の状態を示す目的変数値と、前記除外したサンプルの生

体の状態を示す目的変数値との誤差の代表値を求め、交差検証成績の指標として当該誤差を用いることを特徴とする請求項 23 又は 24 に記載のデータ解析方法。

26. 前記関数が交差検証成績であることを特徴とする請求項 17～25 のいずれかに記載のデータ解析方法。

5 27. 前記関数が交差検証成績と選択された説明変数の個数との関数であることを特徴とする請求項 17～25 のいずれかに記載のデータ解析方法。

28. 前記決定ステップにおいて、少なくとも交差検証成績を独立変数として持つ関数を改善しながら評価判定を繰り返すことを特徴とする請求項 21 に記載のデータ解析方法。

10 29. 前記(1)の選択ステップと前記(2)の計算ステップとを複数のコンピュータで実行させることを特徴とする請求項 17～28 のいずれかに記載のデータ解析方法。

30. 請求項 17、18、19 又は 20 で決定された相関モデル及び予測対象のサンプルについて当該モデルにおいて採用された説明変数を入力する入力ステップと、入力された該説明変数に基づいて該サンプルの生体の状態を予測判定する予測判定ステップからなることを特徴とするデータ解析方法。

15 31. 生体の状態を名義尺度、順序尺度或いは連続量で表現する目的変数とする請求項 18 に記載のデータ解析方法。

32. 前記の統計的手法又は多変量解析手法が、比例ハザード法又はパラメトリックな分布にあてはめた回帰分析法によるモデルを構築する最終モデル決定ステップとからなることを特徴とする請求項 18 又は 20 に記載のデータ解析方法。

20 33. 生体の状態または時間とともに確率的に発生する生体の状態の変化を目的変数とし、複数の遺伝子発現の量および／または細胞内物質の量を説明変数とする相関モデルを決定する、コンピュータにより実行されるデータ解析プログラムであって、

25 生体の状態或いはそれを導出するデータまたは時間とともに確率的に発生する生体の状態の変化に関するデータと、複数の遺伝子発現の量および／または細胞内物質の量からなるサンプルの集合を入力する入力ステップと、

(1)説明変数を選択する選択ステップと、

(2) 部分最小自乗法を実行して交差検証成績を計算する計算ステップまたは前記生体の状態の変化に関するデータにカプラン・マイヤー法又はカトラー・エデラー法による生命表を適用して変化の発生しなかったものの確率を計算して得られた確率を、仮定した分布に基づいた変換または仮定を前提としない変換をし、該変換結果を目的変数とする部分最小自乗法を実行して交差検証成績を計算する計算ステップと、

(3) 前記(2)の計算ステップの結果を評価し、説明変数の採用、不採用を判定する評価判定ステップとを有し、

(4) 前記(1)の選択ステップと前記(2)の計算ステップと前記(3)の評価判定ステップとを実行して部分最小自乗法モデルの少なくとも交差検証成績を独立変数として持つ関数を改善し続けて部分最小自乗法モデルを決定する決定ステップとからなることを特徴とするデータ解析プログラム。

34. 目的変数が生体の状態であって、前記入カステップで入力するデータが生体の状態或いはそれを導出するデータであって、前記(2)の計算ステップが部分最小自乗法を実行して交差検証成績を計算する計算ステップであることを特徴とする請求項33に記載のデータ解析プログラム。

35. 目的変数が時間とともに確率的に発生する生体の状態の変化であって、前記入カステップで入力するデータが時間とともに確率的に発生する生体の状態の変化に関するデータであって、前記(2)の計算ステップが前記生体の状態の変化に関するデータにカプラン・マイヤー法又はカトラー・エデラー法による生命表を適用して変化の発生しなかったものの確率を計算して得られた確率を、仮定した分布に基づいた変換または仮定を前提としない変換をし、該変換結果を目的変数とする部分最小自乗法を実行して交差検証成績を計算する計算ステップであることを特徴とする請求項33に記載のデータ解析プログラム。

36. さらに、前記の決定ステップにて決定された部分最小自乗法モデルに採用されている説明変数又は該モデルの潜在変数を用い、統計的手法又は多変量解析手法によるモデルを構築する最終モデル決定ステップを備えることを特徴とする請求項33、34又は35に記載のデータ解析プログラム。

37. 前記の選択ステップにおいて、説明変数を逐次取捨選択することを特徴

とする請求項 33～36 のいずれかに記載のデータ解析プログラム。

38. 前記の選択ステップにおいて、遺伝的アルゴリズムを用いて説明変数を選択することを特徴とする請求項 33～36 のいずれかに記載のデータ解析プログラム。

5 39. 前記の計算ステップにおいて、1 個のサンプルを逐次除外して部分最小自乗法を実行して交差検証成績を計算することを特徴とする請求項 33～38 のいずれかに記載のデータ解析プログラム。

10 40. 前記の計算ステップにおいて、複数のサンプルを逐次除外して部分最小自乗法を実行して交差検証成績を計算することを特徴とする請求項 33～38 のいずれかに記載のデータ解析プログラム。

15 41. 前記計算ステップにおいて、各計算において除外したサンプルの遺伝子発現から予測される生体の状態を示す目的変数値と、前記除外したサンプルの生体の状態を示す目的変数値との誤差の代表値を求め、交差検証成績の指標として当該誤差を用いることを特徴とする請求項 39 又は 40 に記載のデータ解析プログラム。

42. 前記関数が交差検証成績であることを特徴とする請求項 33～41 のいずれかに記載のデータ解析プログラム。

43. 前記関数が交差検証成績と選択された説明変数の個数との関数であることを特徴とする請求項 33～41 のいずれかに記載のデータ解析プログラム。

20 44. 前記決定ステップにおいて、少なくとも交差検証成績を独立変数として持つ関数を改善しながら評価判定を繰り返すことを特徴とする請求項 37 に記載のデータ解析プログラム。

25 45. 前記(1)の選択ステップと前記(2)の計算ステップとを複数のコンピュータで実行させることを特徴とする請求項 33～44 のいずれかに記載のデータ解析プログラム。

46. 請求項 33、34、35 又は 36 で決定された相関モデル及び予測対象のサンプルについて当該モデルにおいて採用された説明変数を入力する入力ステップと、入力された該説明変数に基づいて該サンプルの生体の状態を予測判定する予測判定ステップからなることを特徴とするデータ解析プログラム。

47. 生体の状態を名義尺度、順序尺度或いは連続量で表現する目的変数とする請求項34に記載のデータ解析プログラム。

48. 前記の統計的手法又は多変量解析手法が、比例ハザード法又はパラメトリックな分布にあてはめた回帰分析法によるモデルを構築する最終モデル決定ステップとからなることを特徴とする請求項34又は36に記載のデータ解析プログラム。

49. 前記の説明変数の選択において、初期状態では説明変数を全く含まないことを特徴とする請求項37に記載のプログラム。

50. 前記の説明変数の選択において、初期状態では全説明変数を含むことを特徴とする請求項37に記載のプログラム。

51. 前記の生体の状態が病気のタイプをあらわす測定値、病気の重篤度をあらわす測定値、病気のタイプをあらわす医療診断の結果、病気の重篤度をあらわす医療診断の結果、あるいはそれらを2次加工した数値であることを特徴とする請求項37～50のいずれかに記載のプログラム。

52. 請求項33～請求項48のいずれかに記載されたプログラムを記録した、コンピュータにより読み取り可能な記録媒体。

53. 実質的にジーンバンクアクセッション番号がU15085、M23452、X52479、U70426、H57330及びS69790からなる遺伝子群の発現を検出することを特徴とするびまん性大細胞型Bリンパ腫の重篤度検定用の細胞内物質測定機材および測定方法並びにびまん性大細胞型Bリンパ腫の重篤度検定方法。

54. さらにジーンバンクアクセッション番号がU03398、M65066、AK001546、BC003536、X00437、U12979、H96306、AA830781及びAA804793からなる群から選択される少なくとも一つの遺伝子の発現を検出することを特徴とする請求項53に記載のびまん性大細胞型Bリンパ腫の重篤度検定用の細胞内物質測定機材および測定方法並びにびまん性大細胞型Bリンパ腫の重篤度検定方法。

55. 実質的にジーンバンクアクセッション番号がAA598572、AA703058及びAA453345からなる遺伝子産物を含む細胞内物質を検出することを特徴とする乳癌の重篤度検定用の細胞内物質測定機材および測定方法並びに乳癌の重篤度検定方法。

56. さらにジーンバンクアクセッション番号がAA406242、H73335、W84753、N71160、AA054669、N32820及びR05667からなる群から選択される少なくとも一つの遺伝子産物を含む細胞内物質を検出することを特徴とする請求項55に記載の乳癌の重篤度検定用の細胞内物質測定機材および測定方法並びに乳癌の重篤度検定方法。

57. 実質的にジーンバンクアクセッション番号がW84753、H08581、AA045730及びAI250654からなる遺伝子産物を含む細胞内物質を検出することを特徴とする乳癌の再発性検定用の細胞内物質測定機材および測定方法並びに乳癌の再発性検定方法。

58. さらにジーンバンクアクセッション番号がAA448641、R78516、R05934、AA629838及びH53037からなる群から選択される少なくとも一つの遺伝子産物を含む細胞内物質を検出することを特徴とする請求項57に記載の乳癌の再発性検定用の細胞内物質測定機材および測定方法並びに乳癌の再発性検定方法。

59. 実質的にジーンバンクアクセッション番号がAA434397、T83209、N53427、N29639、AA485739、AA425861、H84871、T64312、T59518及びAA037488からなる遺伝子産物を含む細胞内物質を検出することを特徴とする乳癌の再発性検定用の細胞内物質測定機材および測定方法並びに乳癌の再発性検定方法。

60. さらにジーンバンクアクセッション番号がAA406231の遺伝子産物を含む細胞内物質を検出することを特徴とする請求項59に記載の乳癌の再発性検定用の細胞内物質測定機材および測定方法並びに乳癌の再発性検定方法。

61. 実質的にジーンバンクアクセッション番号がH11482、T64312及びAA045340からなる遺伝子産物を含む細胞内物質を検出することを特徴とする乳癌の再発性検定用の細胞内物質測定機材および測定方法並びに乳癌の再発性検定方法。

図 1

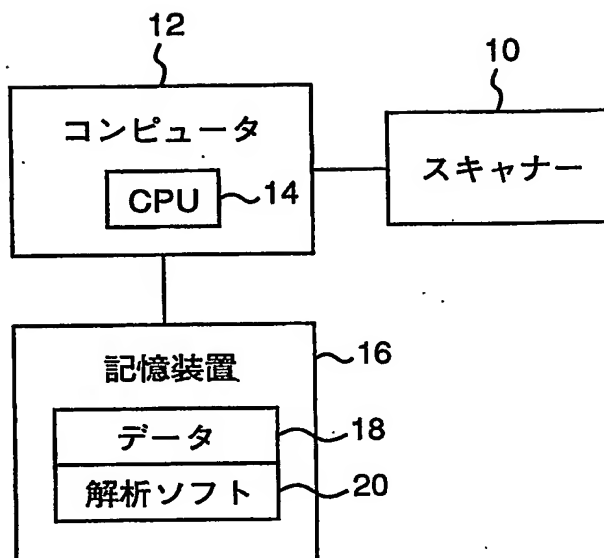


図 2

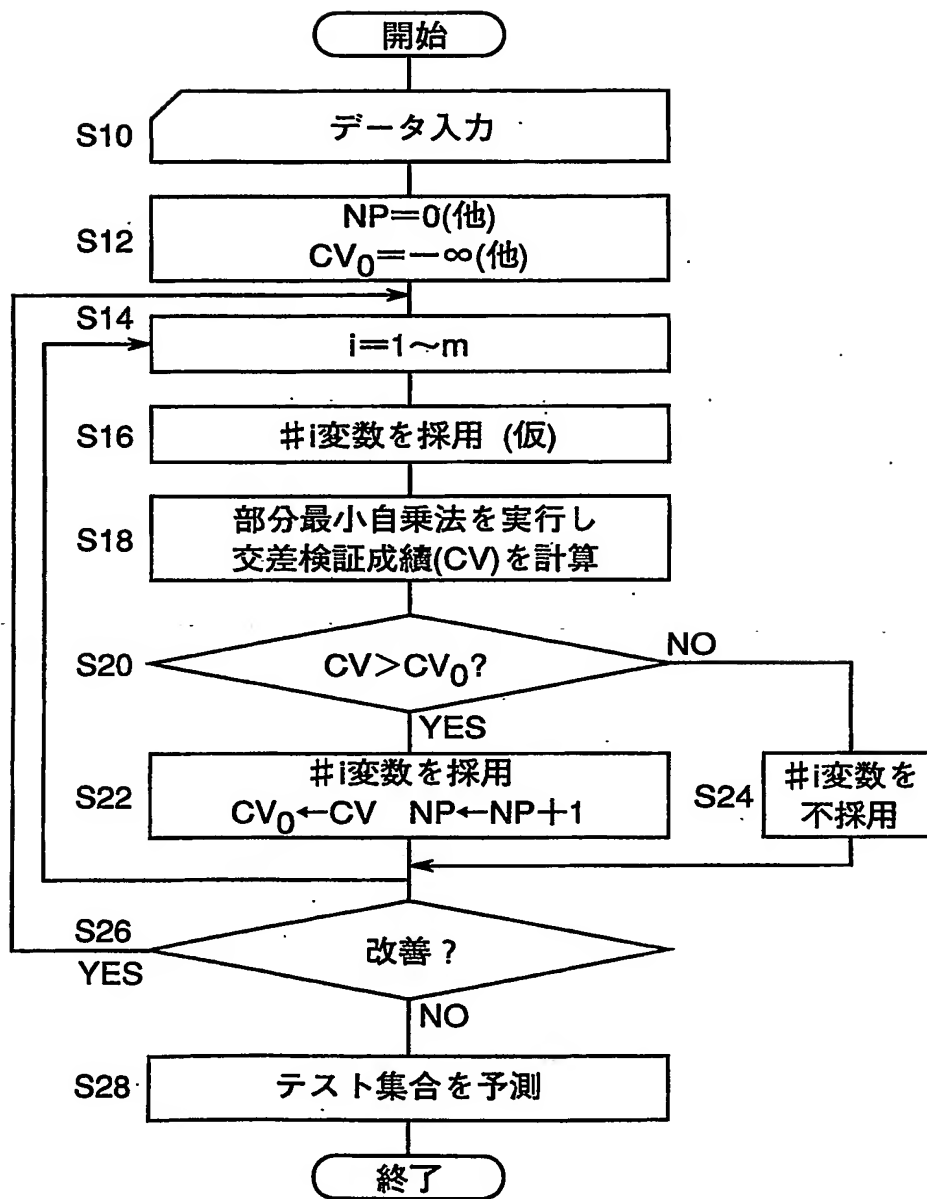


図 3

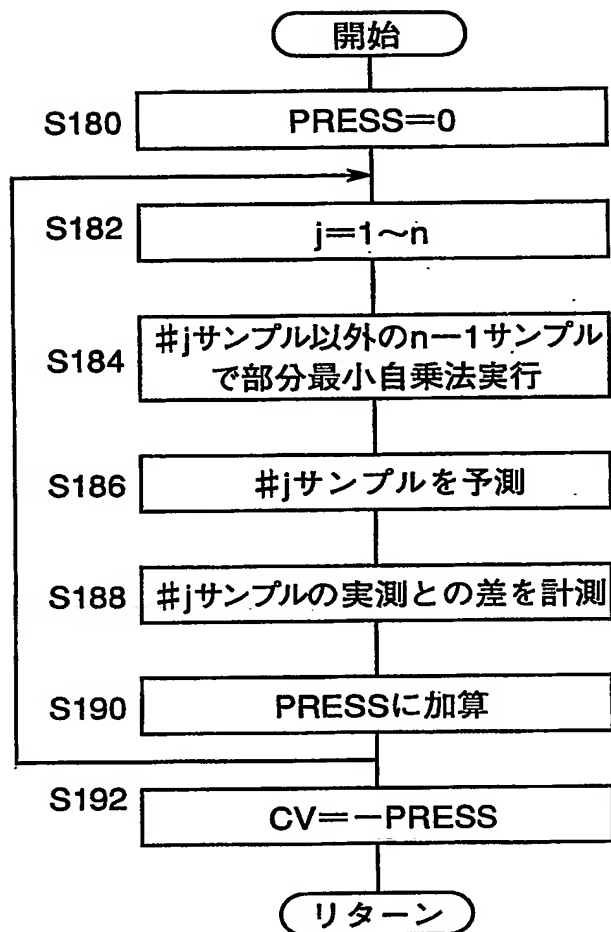


図 4

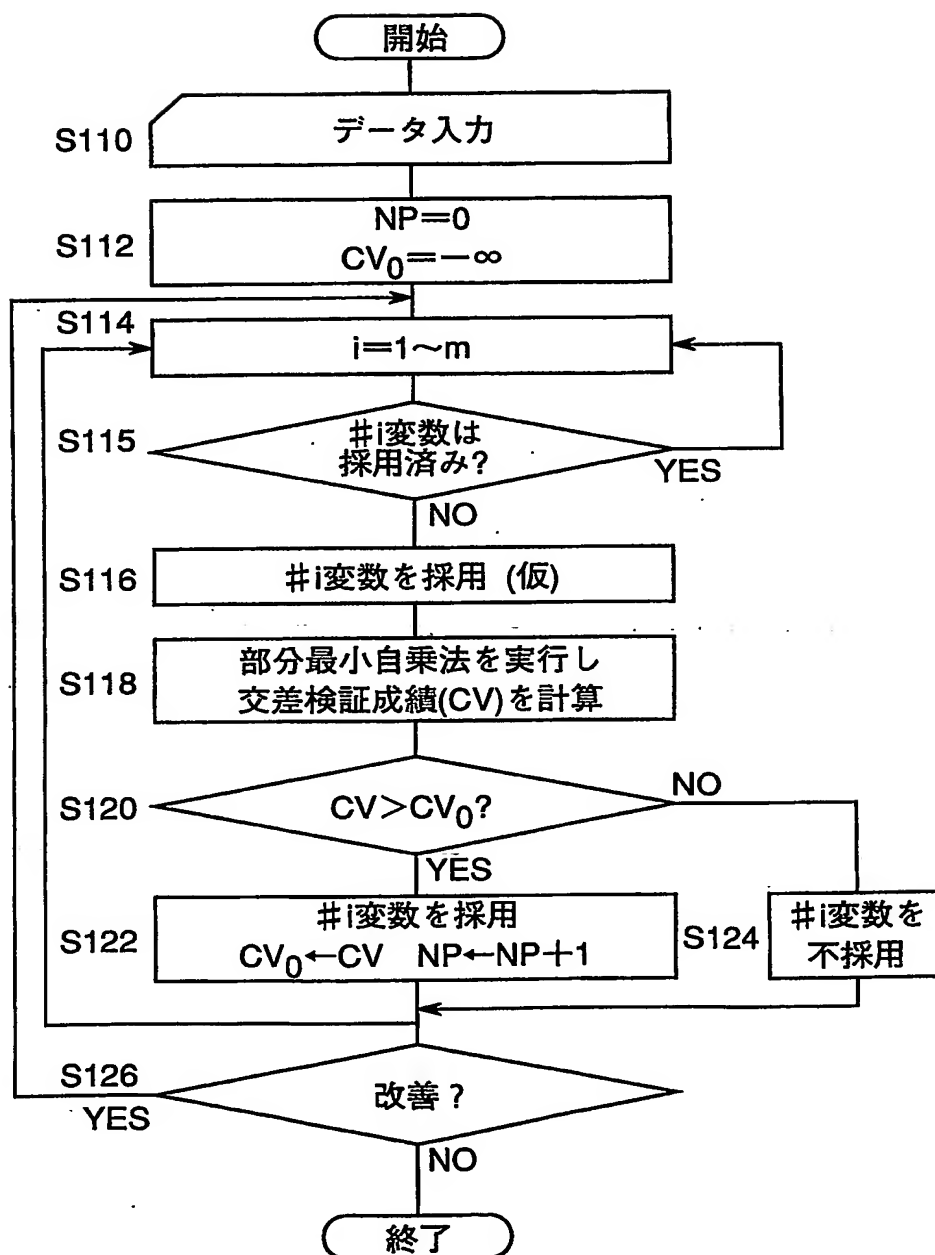


図 5

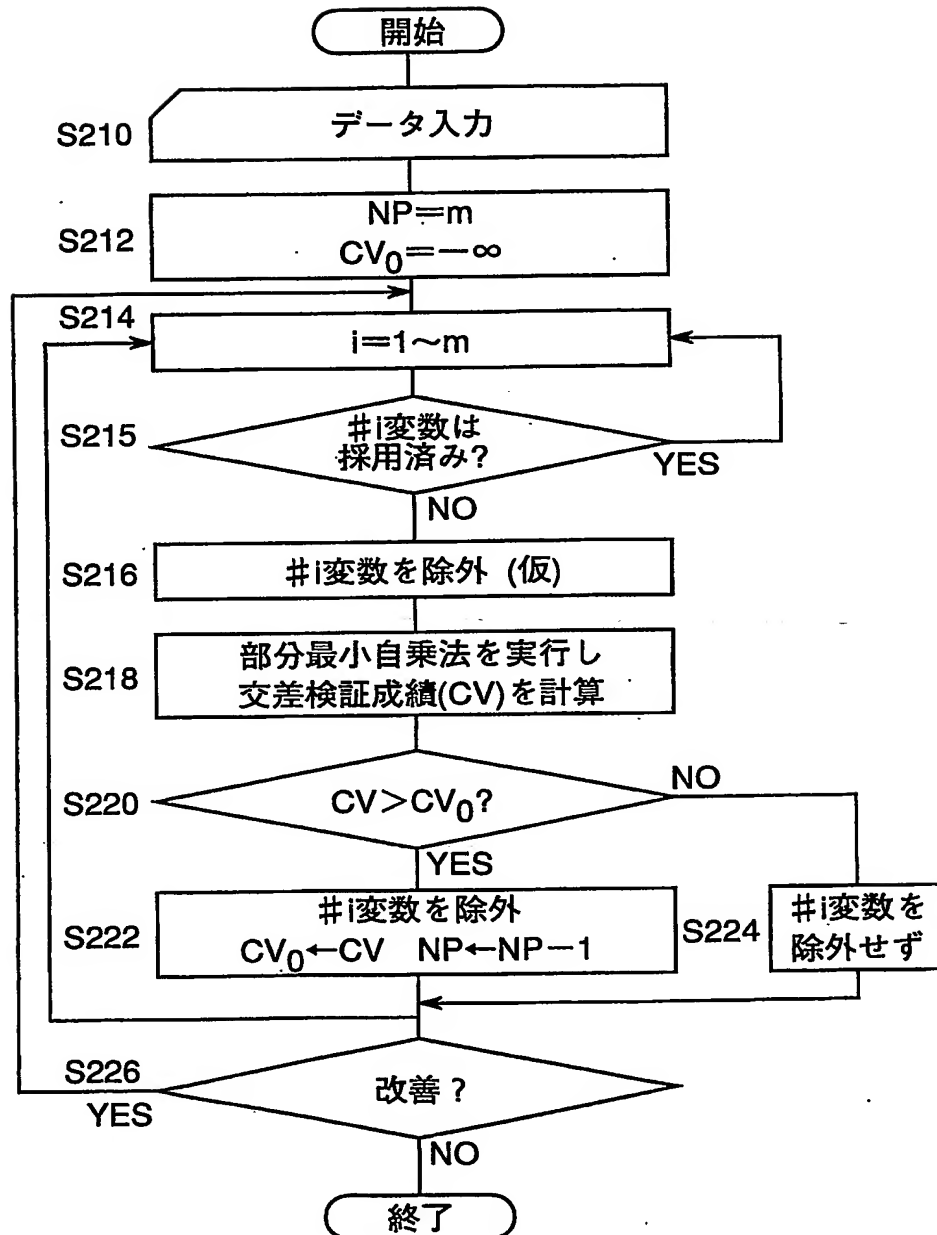


図 6

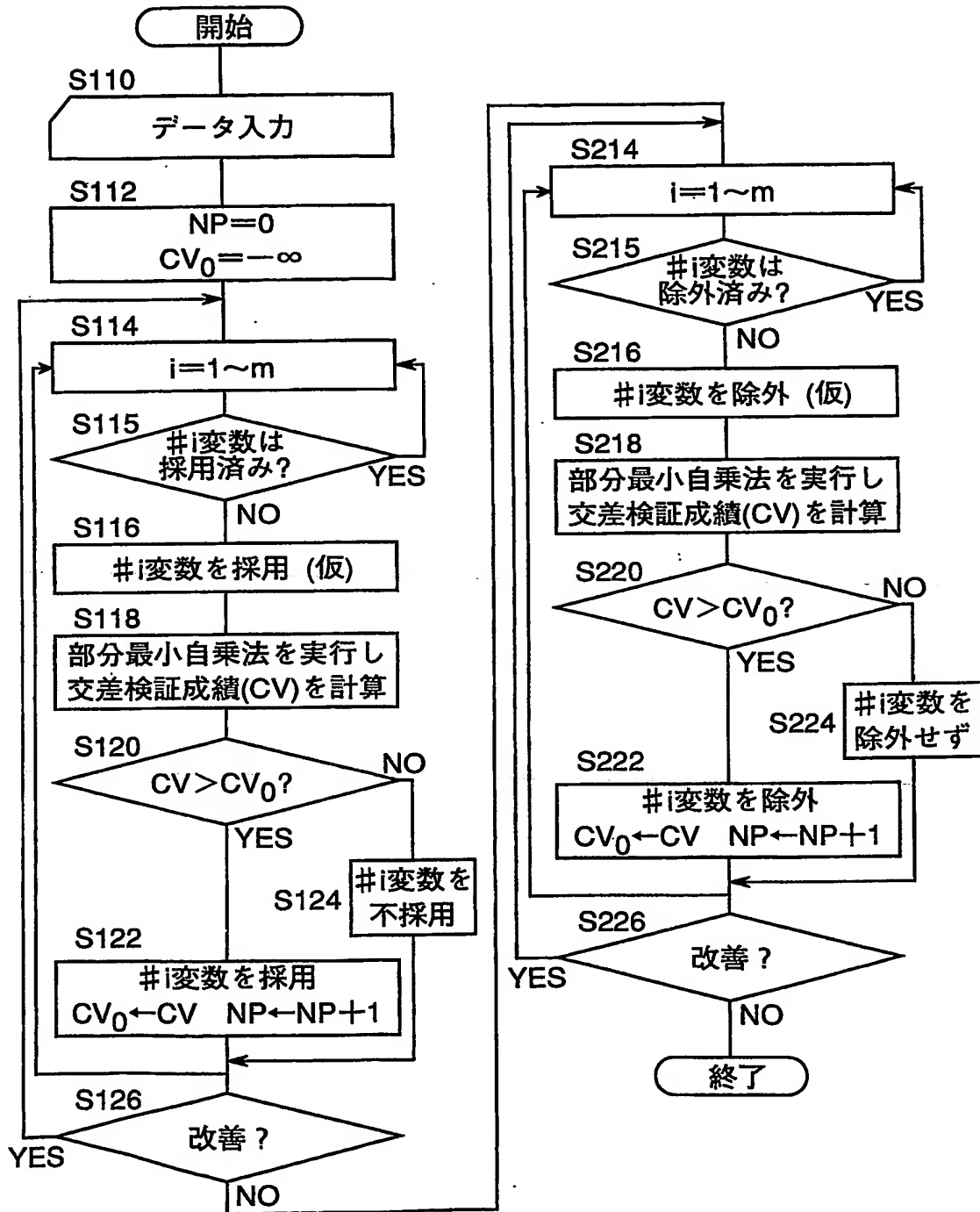


図7

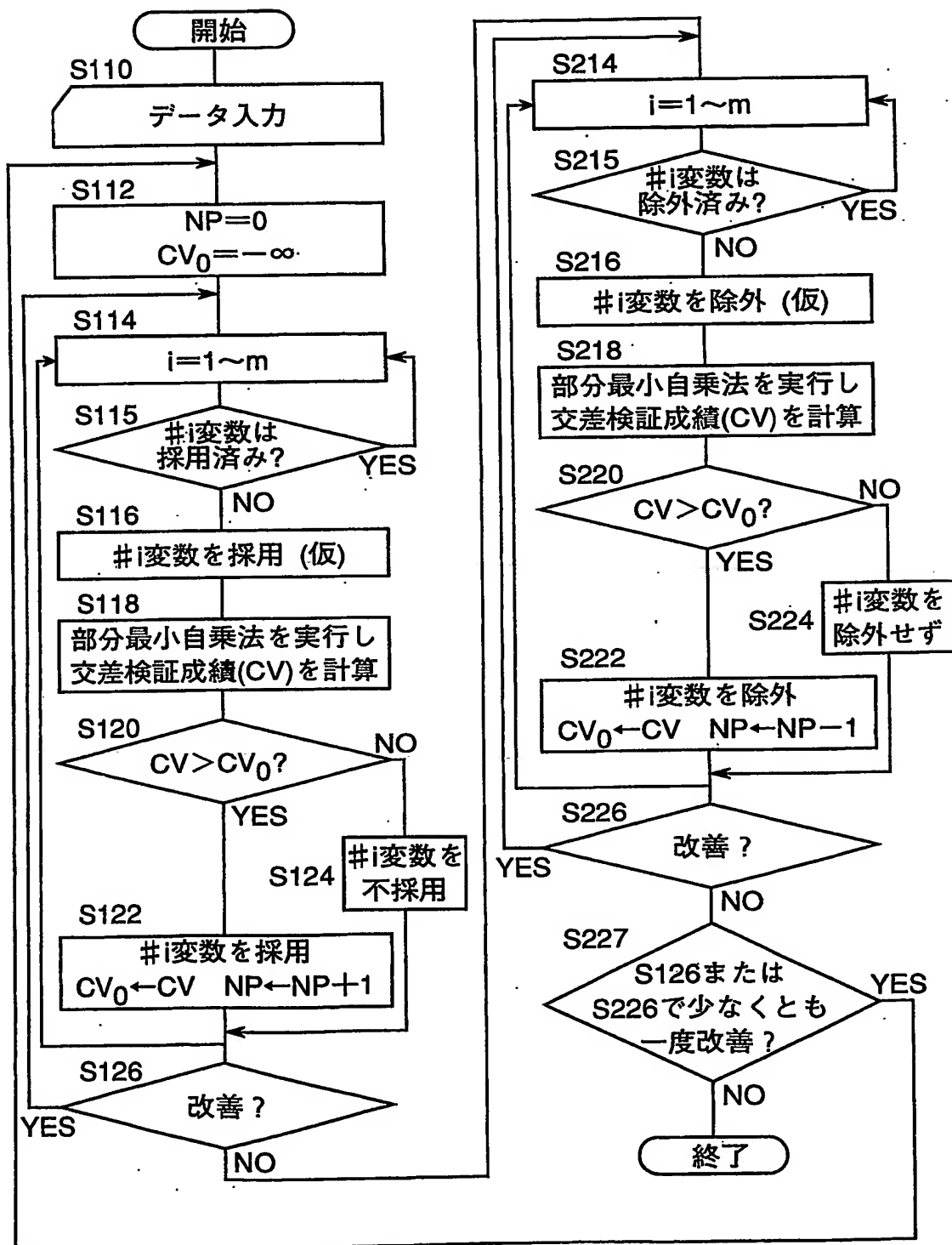


図 8

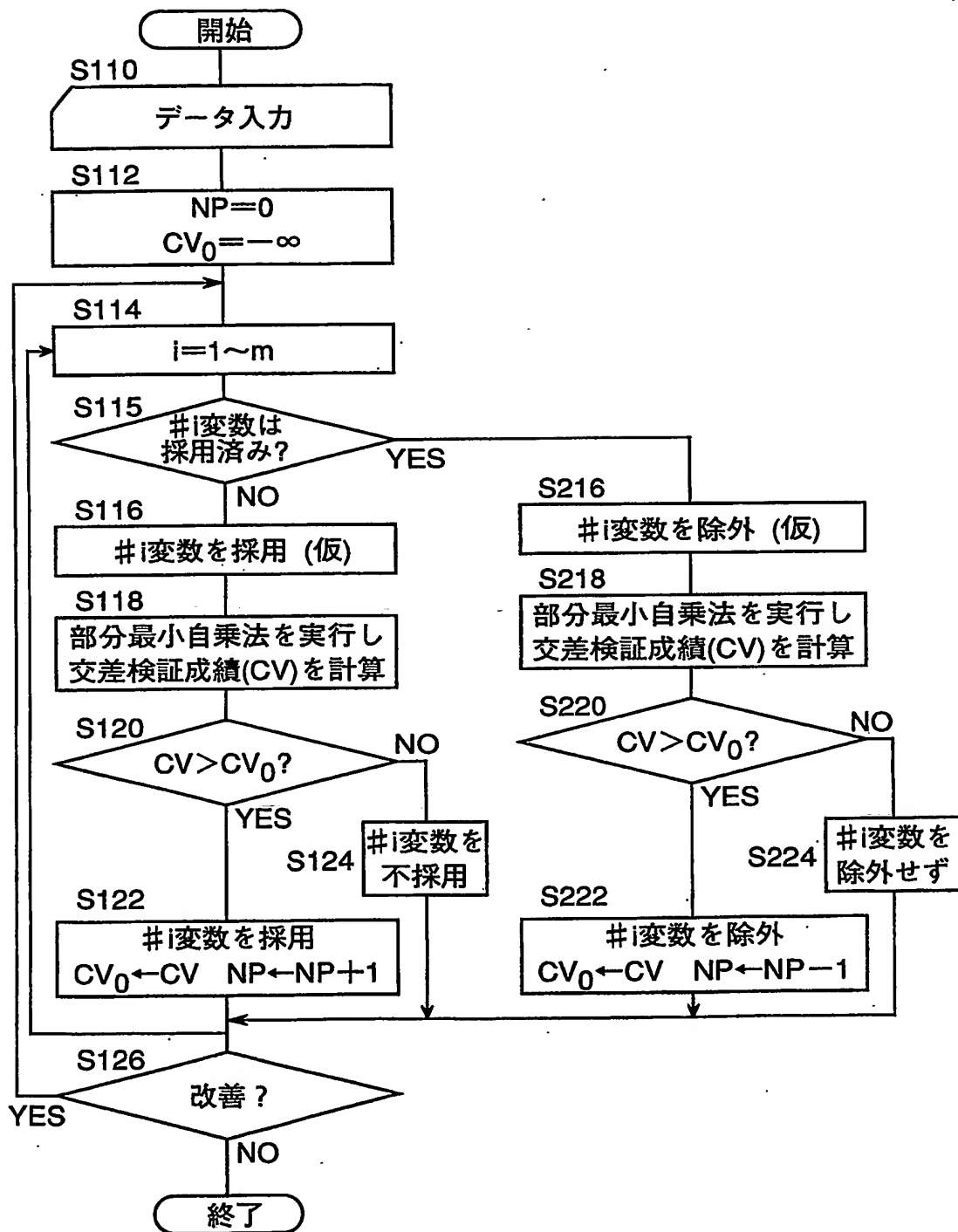


図 9

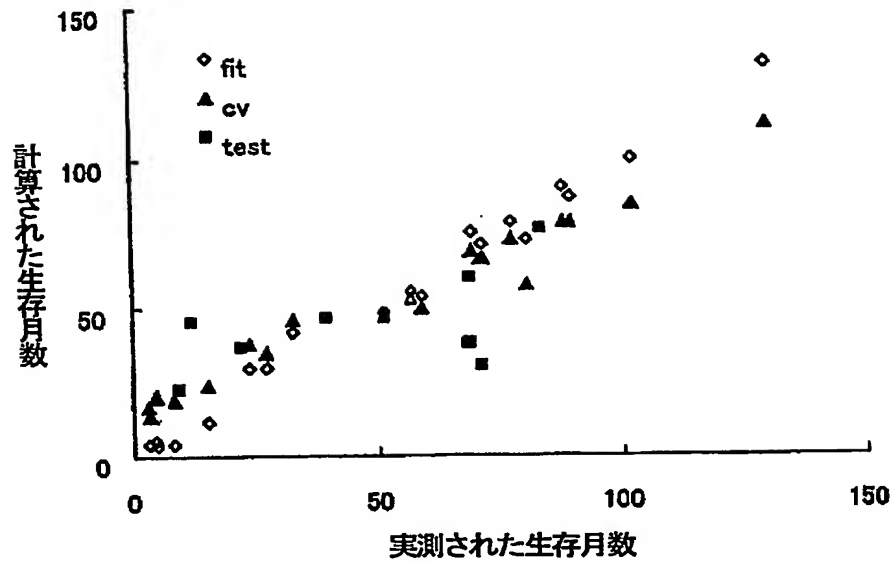


図 10

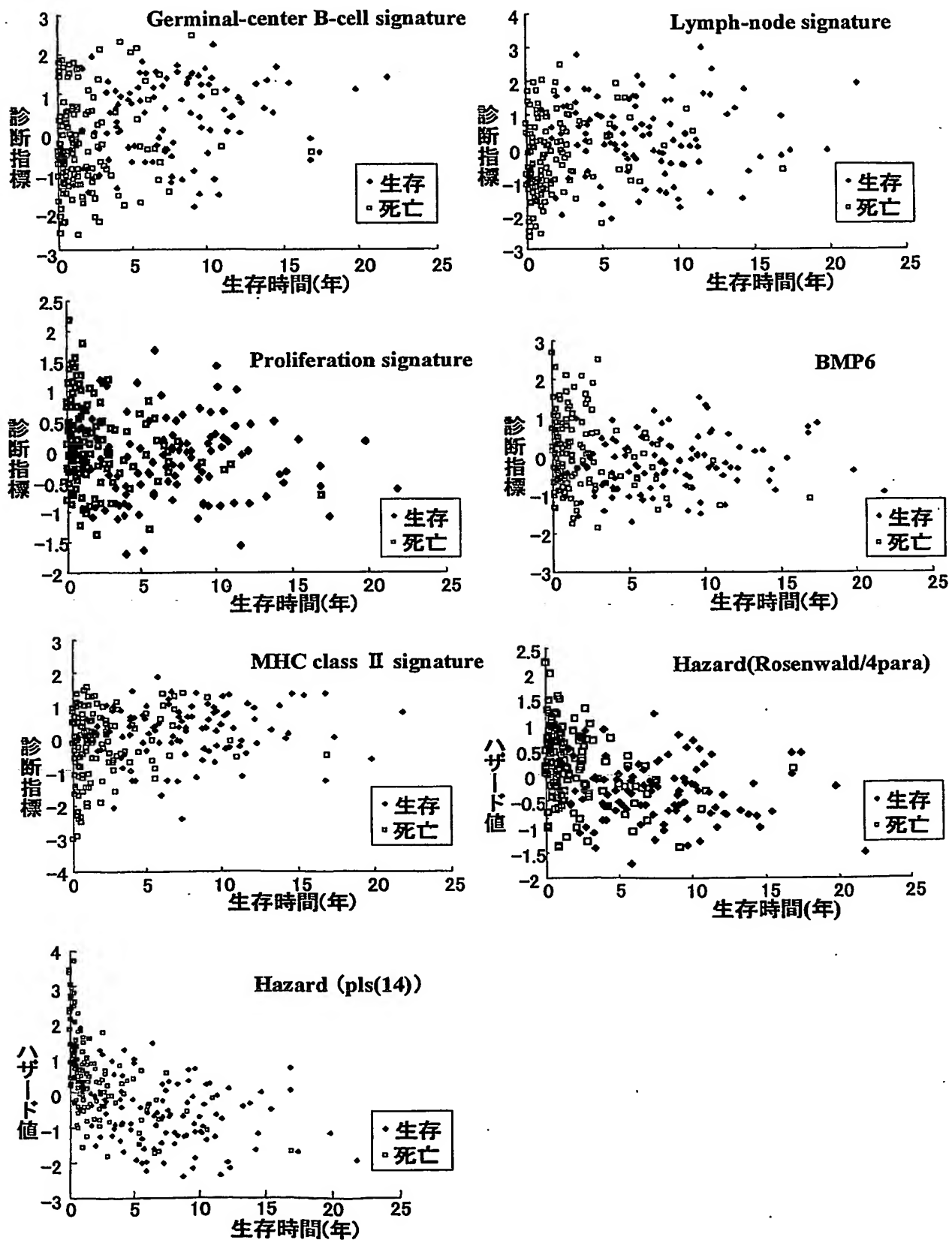


図 1 1

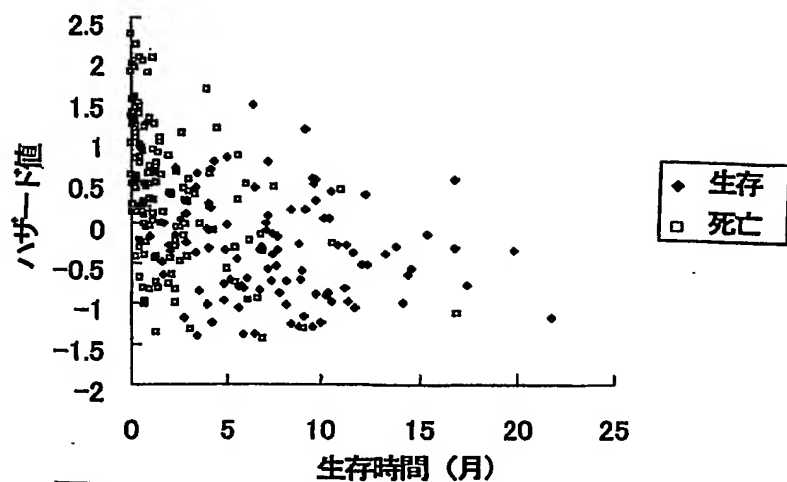


図 1 2

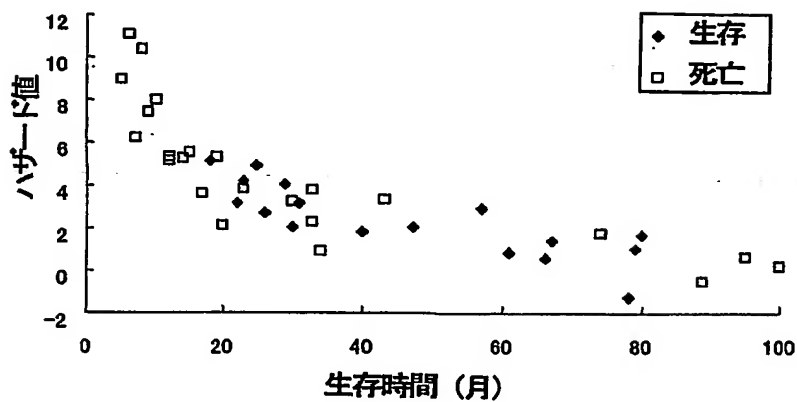
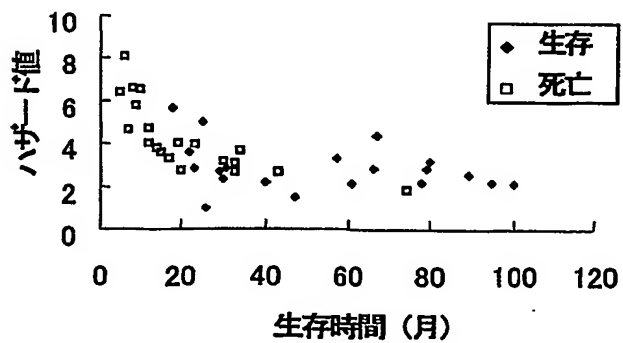


図 1 3



12/14

図 1.4

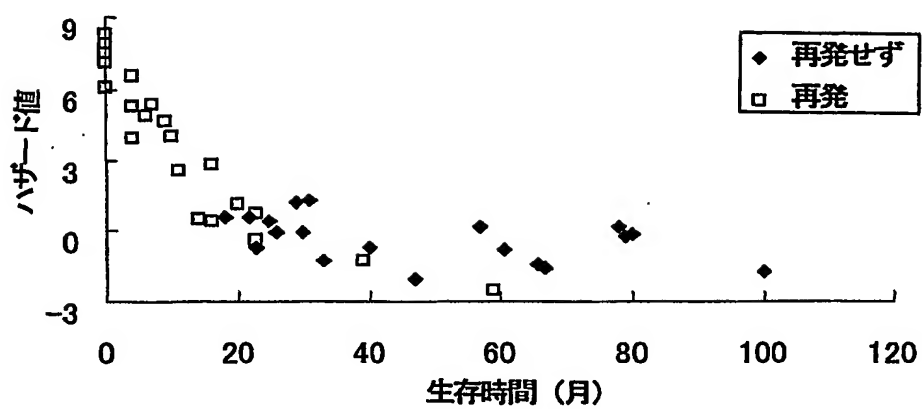


図 1.5

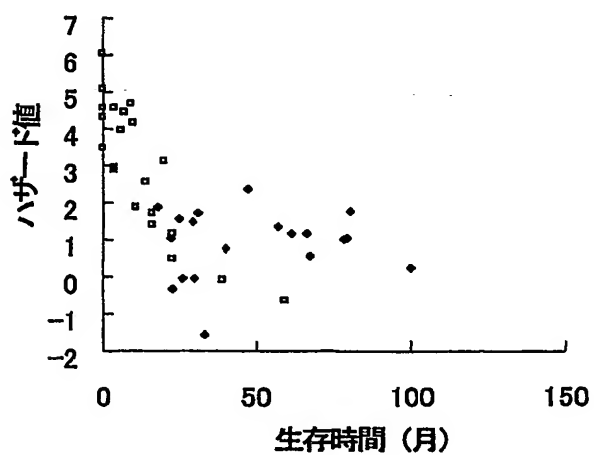


図 1 6

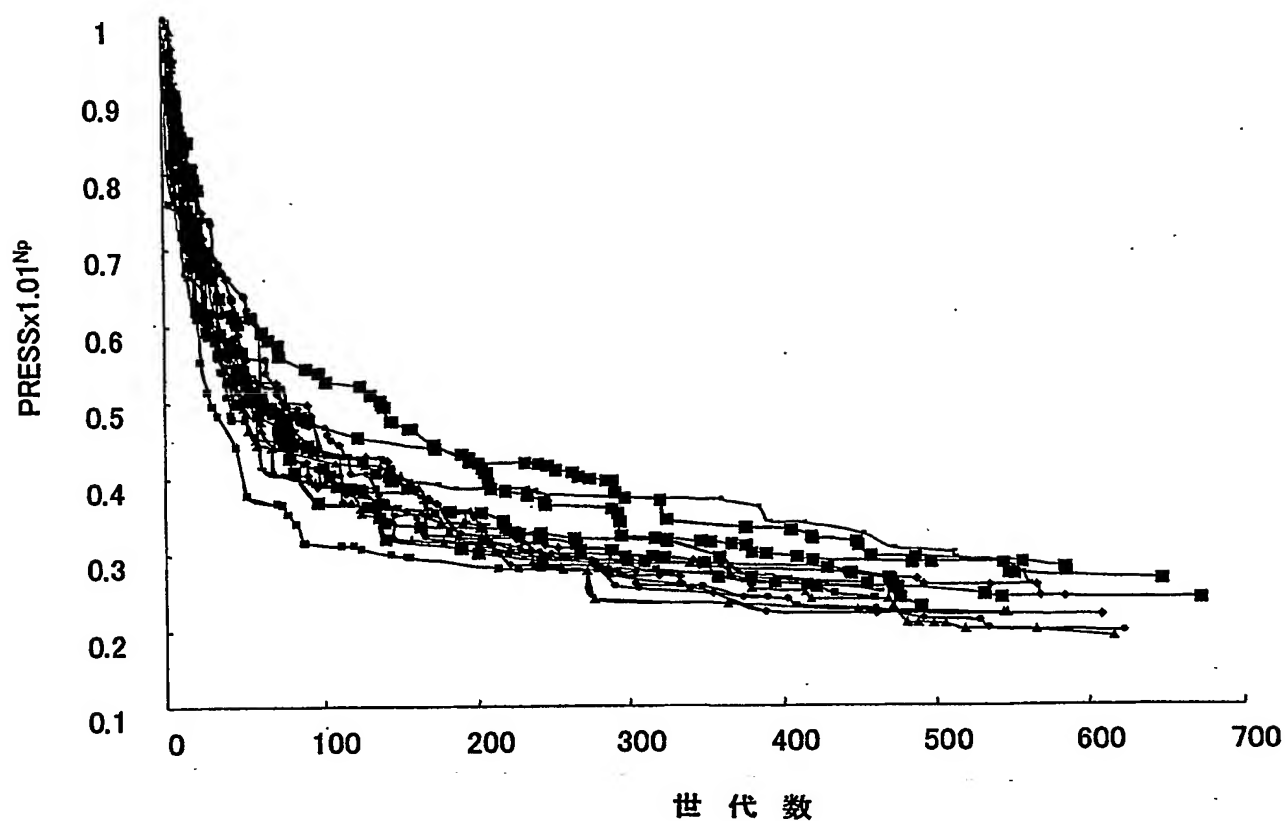


図 1 7

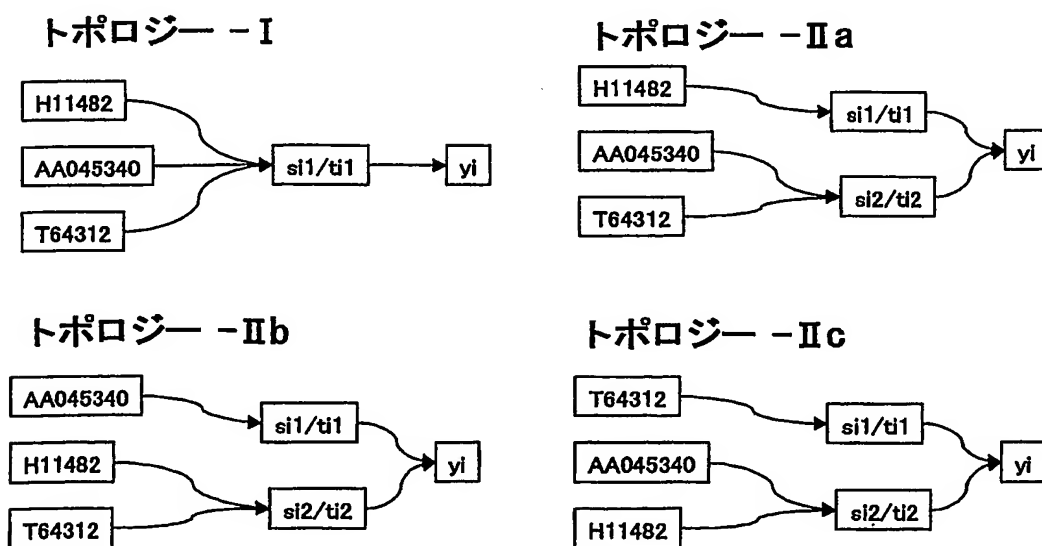


図 18

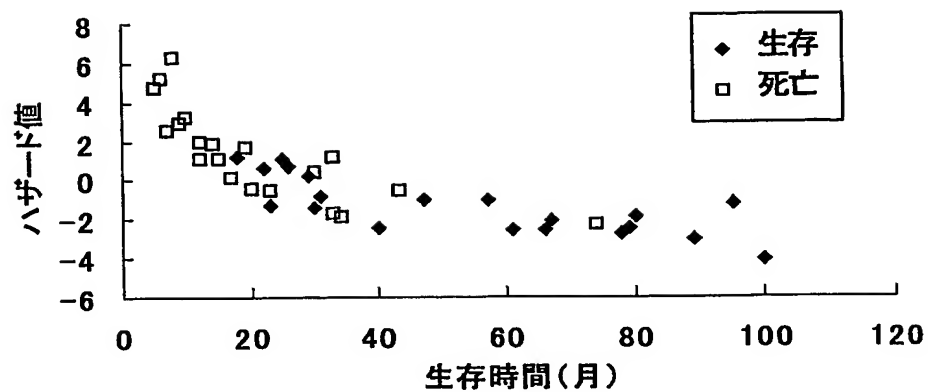
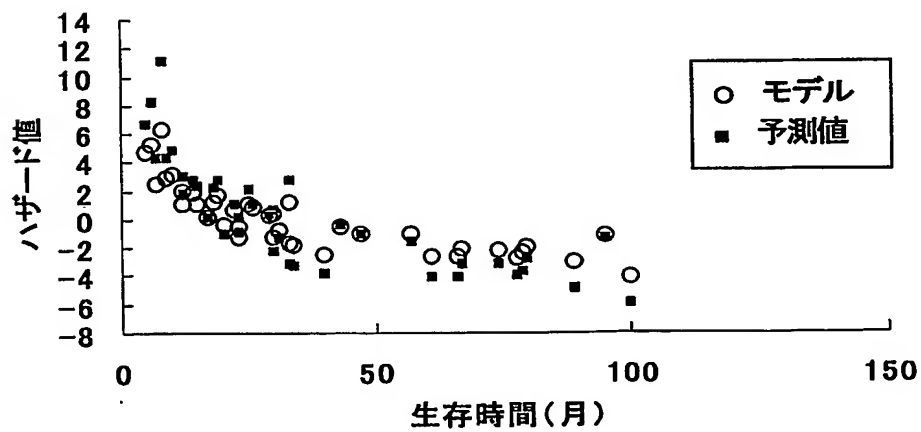


図 19



INTERNATIONAL SEARCH REPORT

International Application No.

PCT/JP03/04059

A. CLASSIFICATION OF SUBJECT MATTER Int.Cl ⁷ G06F17/18, G06F17/30, C12N15/00, C12Q1/68, G01N33/574				
According to International Patent Classification (IPC) or to both national classification and IPC				
B. FIELDS SEARCHED Minimum documentation searched (classification system followed by classification symbols) Int.Cl ⁷ G06F17/00-17/18, G06F17/30, C12N15/00-15/90, C12Q1/00-1/70, G01N33/48-33/98				
Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched Jitsuyo Shinan Koho 1922-1996 Toroku Jitsuyo Shinan Koho 1994-2003 Kokai Jitsuyo Shinan Koho 1971-2003 Jitsuyo Shinan Toroku Koho 1996-2003				
Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)				
C. DOCUMENTS CONSIDERED TO BE RELEVANT				
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.		
A	WO 02/25405 A2 (THE REGENTS OF THE UNIVERSITY OF CALIFORNIA), 28 March, 2002 (28.03.02), Full text; all drawings (Family: none)	1-52		
A	WO 00/70340 A2 (KAROLINSKA INNOVATIONS AB.), 23 November, 2000 (23.11.00), Full text; all drawings & EP 1179175 A3	1-52		
A	P. Gramatica, et al., QSAR STUDY ON THE TROPOSPHERIC DEGRADATION OF ORGANIC COMPOUNDS. In: Chemosphere 1999, Vol.38, No.6, pages 1371 to 1378	1-52		
<input checked="" type="checkbox"/> Further documents are listed in the continuation of Box C. <input type="checkbox"/> See patent family annex.				
<table border="0"> <tr> <td> * Special categories of cited documents: "A" document defining the general state of the art which is not considered to be of particular relevance "E" earlier document but published on or after the international filing date "L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified) "O" document referring to an oral disclosure, use, exhibition or other means "P" document published prior to the international filing date but later than the priority date claimed </td> <td> "T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention "X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone "Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art "&" document member of the same patent family </td> </tr> </table>			* Special categories of cited documents: "A" document defining the general state of the art which is not considered to be of particular relevance "E" earlier document but published on or after the international filing date "L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified) "O" document referring to an oral disclosure, use, exhibition or other means "P" document published prior to the international filing date but later than the priority date claimed	"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention "X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone "Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art "&" document member of the same patent family
* Special categories of cited documents: "A" document defining the general state of the art which is not considered to be of particular relevance "E" earlier document but published on or after the international filing date "L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified) "O" document referring to an oral disclosure, use, exhibition or other means "P" document published prior to the international filing date but later than the priority date claimed	"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention "X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone "Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art "&" document member of the same patent family			
Date of the actual completion of the international search 27 June, 2003 (27.06.03)		Date of mailing of the international search report 08 July, 2003 (08.07.03)		
Name and mailing address of the ISA/ Japanese Patent Office		Authorized officer		
Facsimile No.		Telephone No.		

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP03/04059

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	J. Khan et al., Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. In: NATURE MEDICINE, 2001, Vol.7, No.6, pages 673 to 679	1-52
P,A	A. ROSENWALD et al., THE USE OF MOLECULAR PROFILING TO PREDICT SURVIVAL AFTER CHEMOTHERAPY FOR DIFFUSE LARGE-B-CELL LYMPHOMA. In: The New England Journal of Medicine, 20 June, 2002 (20.06.02), Vol.346, No.25, pages 1937 to 1947	53-61

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP03/04059

Box I Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box II Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

The claimed inventions in the present application consist of 6 groups of inventions, i.e., the inventions according to claims 1 to 52, the inventions according to claims 53 and 54, the inventions according to claims 55 and 56, the inventions according to claims 57 and 58, the inventions according to 59 and 60 and the invention according to claim 61.

As reported in, for example, International Publication WO 00/70340 pamphlet (2000), it had been well known by a skilled person in the art to detect an important gene based on the analysis results of determining a partial least squares method model or the like. Such being the case, there is no (continued to extra sheet)

1. ☒ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying an additional fee, this Authority did not invite payment of any additional fee.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest ☐ The additional search fees were accompanied by the applicant's protest.
☒ No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP03/04059

Continuation of Box No.II of continuation of first sheet(1)

technical relationship involving any "special technical feature" among these 6 groups of inventions and these groups of inventions are not considered as a group of inventions so linked as to form a single general inventive concept.

A. 発明の属する分野の分類 (国際特許分類 (IPC))

Int. Cl⁷ G06F17/18, G06F17/30, C12N15/00, C12Q1/68,
G01N33/574

B. 調査を行った分野

調査を行った最小限資料 (国際特許分類 (IPC))

Int. Cl⁷ G06F17/00-17/18, G06F17/30, C12N15/00-15/90,
C12Q1/00-1/70, G01N33/48-33/98

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	1922-1996年
日本国公開実用新案公報	1971-2003年
日本国登録実用新案公報	1994-2003年
日本国実用新案登録公報	1996-2003年

国際調査で使用した電子データベース (データベースの名称、調査に使用した用語)

C. 関連すると認められる文献

引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求の範囲の番号
A	WO 02/25405 A2 (THE REGENTS OF THE UNIVERSITY OF CALIFORNIA) 2002.03.28, 全文, 全図 (ファミリーなし)	1-52
A	WO 00/70340 A2 (KAROLINSKA INNOVATIONS AB) 2000.11.23, 全文, 全図 & EP 1179175 A3	1-52
A	P. Gramatica, 外2名, QSAR STUDY ON THE TROPOSPHERIC DEGRADATION OF ORGANIC COMPOUNDS. In: Chemosphere, 1999, vol.38, no.6, p.1371-1378	1-52

☒ C欄の続きにも文献が列挙されている。☐ パテントファミリーに関する別紙を参照。

* 引用文献のカテゴリー

- 「A」 特に関連のある文献ではなく、一般的技術水準を示すもの
- 「E」 国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの
- 「L」 優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献 (理由を付す)
- 「O」 口頭による開示、使用、展示等に言及する文献
- 「P」 国際出願日前で、かつ優先権の主張の基礎となる出願

の日の後に公表された文献

- 「T」 国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの
- 「X」 特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
- 「Y」 特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
- 「&」 同一パテントファミリー文献

国際調査を完了した日

27.06.03

国際調査報告の発送日

08.07.03

国際調査機関の名称及びあて先

日本国特許庁 (ISA/J P)
郵便番号100-8915
東京都千代田区霞が関三丁目4番3号

特許庁審査官 (権限のある職員)

羽立 章二



5B

2944

電話番号 03-3581-1101 内線 3545

C (続き) . 関連すると認められる文献

引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求の範囲の番号
A	J. Khan et. al., Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. In: NATURE MEDICINE, 2001, vol. 7, no. 6, p. 673-679	1-52
P, A	A. ROSENWALD et. al., THE USE OF MOLECULAR PROFILING TO PREDICT SURVIVAL AFTER CHEMOTHERAPY FOR DIFFUSE LARGE-B-CELL LYMPHOMA. In: The New England Journal of Medicine, 2002. 06. 20, vol. 346, no. 25, p. 1937-1947	53-61

第I欄 請求の範囲の一部の調査ができないときの意見 (第1ページの2の続き)

法第8条第3項 (PCT 17条(2)(a)) の規定により、この国際調査報告は次の理由により請求の範囲の一部について作成しなかった。

1. ☐ 請求の範囲 _____ は、この国際調査機関が調査をすることを要しない対象に係るものである。つまり、
2. ☐ 請求の範囲 _____ は、有意義な国際調査をすることができる程度まで所定の要件を満たしていない国際出願の部分に係るものである。つまり、
3. ☐ 請求の範囲 _____ は、従属請求の範囲であってPCT規則6.4(a)の第2文及び第3文の規定に従って記載されていない。

第II欄 発明の単一性が欠如しているときの意見 (第1ページの3の続き)

次に述べるようにこの国際出願に二以上の発明があるとこの国際調査機関は認めた。

本願の請求の範囲に記載された発明は、請求の範囲1乃至52に係る発明、請求の範囲53及び54に係る発明、請求の範囲55及び56に係る発明、請求の範囲57及び58に係る発明、請求の範囲59及び60に係る発明、並びに、請求の範囲61に係る発明という6群の発明からなるものである。

例えば国際公開第00/70340号パンフレット(2000)に記載されているように部分最小自乗法診断モデルを決定する等により解析結果に基づいて重要な遺伝子を検出したことは当業者によく知られている事項であることに鑑みると、これら6群の発明のそれぞれは「特別な技術的特徴」を含む技術的な関係にないから、単一の一般的発明概念を形成するように連関しているものとは認められない。

1. ☒ 出願人が必要な追加調査手数料をすべて期間内に納付したので、この国際調査報告は、すべての調査可能な請求の範囲について作成した。
2. ☐ 追加調査手数料を要求するまでもなく、すべての調査可能な請求の範囲について調査することができたので、追加調査手数料の納付を求めなかった。
3. ☐ 出願人が必要な追加調査手数料を一部のみしか期間内に納付しなかったので、この国際調査報告は、手数料の納付のあった次の請求の範囲のみについて作成した。
4. ☐ 出願人が必要な追加調査手数料を期間内に納付しなかったので、この国際調査報告は、請求の範囲の最初に記載されている発明に係る次の請求の範囲について作成した。

追加調査手数料の異議の申立てに関する注意

- ☐ 追加調査手数料の納付と共に出願人から異議申立てがあった。
☒ 追加調査手数料の納付と共に出願人から異議申立てがなかった。